Stat 5102 Lecture Slides
Deck 2

Charles J. Geyer
School of Statistics
University of Minnesota

Statistical Inference

Statistics is probability done backwards.

In probability theory we give you one probability model, also called a probability distribution. Your job is to say something about expectations, probabilities, quantiles, etc. for that distribution. In short, given a probability model, describe data from that model.

In theoretical statistics, we give you a *statistical model*, which is a family of probability distributions, and we give you some data assumed to have one of the distributions in the model. Your job is to say something about which distribution that is. In short, given a statistical model and data, infer which distribution in the model is the one for the data.

Statistical Models

A *statistical model* is a family of probability distributions.

A *parametric statistical model* is a family of probability distributions specified by a finite set of parameters. Examples: $\text{Ber}(p)$, $\mathcal{N}(\mu, \sigma^2)$, and the like.

A *nonparametric statistical model* is a family of probability distributions too big to be specified by a finite set of parameters. Examples: all probability distributions on $\mathbb{R}$, all continuous symmetric probability distributions on $\mathbb{R}$, all probability distributions on $\mathbb{R}$ having second moments, and the like.

Statistical Models and Submodels

If $\mathcal{M}$ is a statistical model, it is a family of probability distributions.

A *submodel* of a statistical model $\mathcal{M}$ is a family of probability distributions that is a subset of $\mathcal{M}$.

If $\mathcal{M}$ is parametric, then we often specify it by giving the PMF (if the distributions are discrete) or PDF (if the distributions are continuous)

$$\{ f_\theta : \theta \in \Theta \}$$

where Θ is the parameter space of the model.

Statistical Models and Submodels (cont.)

We can have models and submodels for nonparametric families too.

All probability distributions on $\mathbb{R}$ is a statistical model.

All continuous and symmetric probability distributions on $\mathbb{R}$ is a submodel of that.

All univariate normal distributions is a submodel of that.

The first two are nonparametric. The last is parametric.

Statistical Models and Submodels (cont.)

Submodels of parametric families are often specified by fixing the values of some parameters.

All univariate normal distributions is a statistical model.

All univariate normal distributions with known variance is a submodel of that. Its only unknown parameter is the mean. Its parameter space is $\mathbb{R}$.

All univariate normal distributions with known mean is a different submodel. Its only unknown parameter is the variance. Its parameter space is $(0, \infty)$.

Statistical Models and Submodels (cont.)

Thus $\mathcal{N}(\mu, \sigma^2)$ does not, by itself, specify a statistical model. You must say what the parameter space is. Alternatively, you must say which parameters are considered known and which unknown.

The parameter space is the set of all possible values of the unknown parameter.

If there are several unknown parameters, we think of them as components of the unknown parameter vector, the set of all possible values of the unknown parameter vector is the parameter space.

Parameters

The word "parameter" has two closely related meanings in statistics.

- One of a finite set of variables that specifies a probability distribution within a family. Examples: p for Ber(p), and μ and σ^2 for $\mathcal{N}(\mu, \sigma^2)$.

- A numerical quantity that can be specified for all probability distributions in the family. Examples: mean, median, variance, upper quartile.

Parameters (cont.)

The first applies only to parametric statistical models. The parameters are the parameters of the model. The second applies to nonparametric statistical models too.

Every distribution has a median. If it is not unique, take any unique definition, say $G(1/2)$, where G is the quantile function.

Not every distribution has a mean. But if the family in question is all distributions with first moments, then every distribution in the family has a mean.

Truth

The word "true" has a technical meaning in statistics. In the phrase "true unknown parameter" or "true unknown distribution" it refers to the probability distribution of the data, which is assumed (perhaps incorrectly) to be one of the distributions in the statistical model under discussion.

Statistics

The word "statistic" (singular) has a technical meaning in statistics (plural, meaning the subject).

A *statistic* is a function of data only, not parameters. Hence a statistic can be calculated from the data for a problem, even though the true parameter values are unknown.

The sample mean $\overline{X}_n$ is a statistic, so is the sample variance S_n^2, and so is the sample median $\widetilde{X}_n$.

Statistics (cont.)

All scalar-valued statistics are random variables, but not all random variables are statistics. Example: $(\overline{X}_n - \mu)/(S_n/\sqrt{n})$ is a random variable but not a statistic, because it contains the parameter μ.

Statistics can also be random vectors. Example: $(\overline{X}_n, S_n^2)$ is a two-dimensional vector statistic.

Estimates

A statistic X is an *estimate* of the parameter θ if we say so.

The term "estimate" does not indicate that X has any particular properties. It only indicates our intention to use X to say something about the true unknown value of the parameter θ.

There can be many different estimates of a parameter θ. The sample mean $\overline{X}_n$ is an obvious estimate of the population mean μ. The sample median $\widetilde{X}_n$ is a less obvious estimate of μ. The sample standard deviation S_n is a silly estimate of μ. The constant random variable X always equal to 42 is another a silly estimate of μ.

Estimates (cont.)

We often indicate the connection between a statistic and the parameter it estimates by putting a hat on the parameter. If θ is a parameter, we denote the statistic $\widehat{\theta}$ or $\widehat{\theta}_n$ if we also want to indicate the sample size.

The formal name for the symbol $\widehat{\ }$ is "caret" but statisticians always say "hat" and read $\widehat{\theta}$ as "theta hat".

Estimates (cont.)

The conventions are now getting a bit confusing.

Capital lightface roman letters like X, Y, Z denote statistics.

Sometimes they are decorated by bars, wiggles, and subscripts, like $\overline{X}_n$ and $\widetilde{X}_n$, but they are still statistics.

Parameters are denoted by greek letters like μ, σ, and θ, and, of course, any function of a parameter is a parameter, like σ^2.

Exception: we and nearly everybody else use p for the parameter of the Ber(p), Bin(n, p), Geo(p), and NegBin(n, p) distributions, perhaps because the greek letter with the "p" sound is π and it is a frozen letter that always means the number 3.1415926535..., so we can't use that.

Estimates (cont.)

Whatever the reason for the exception, we do have the convention roman letters for statistics and greek letters for parameters except for the exceptions.

Now we have a different convention. Greek letters with hats are statistics not parameters.

θ is the parameter that the statistic $\widehat{\theta}_n$ estimates.

μ is the parameter that the statistic $\widehat{\mu}_n$ estimates.

Theories of Statistics

There is more than one way to do statistics. We will learn two ways: *frequentist* and *Bayesian*. There are other theories, but we won't touch them.

Frequentist Statistics

The frequentist theory of probability can only define probability for an infinite sequence of IID random variables X_1, X_2, ... It defines the probability $\Pr(X_i \in A)$, which is the same for all i because the X_i are identically distributed, as what the corresponding expectation for the empirical distribution

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(X_i)$$

converges to. We know

$$P_n(A) \xrightarrow{P} \Pr(X_i \in A)$$

by the LLN. But the frequentist theory tries to turn this into a definition rather than a theorem.

Frequentist Statistics (cont.)

The frequentist theory of probability has some appeal to philosophers but no appeal to mathematicians. The attempt to make a theorem so complicated we didn't even prove it a fundamental definition makes the frequentist theory so difficult that no one uses it.

That's why everyone uses the formalist theory: if we call it a probability and it obeys the axioms for probability (5101, deck 2, slides 2–4 and 132–141), then it is a probability.

Frequentist Statistics (cont.)

The frequentist theory of statistics is completely different from the frequentist theory of probability.

The frequentist theory of statistics uses sampling distributions. If $\widehat{\theta}_n$ is an estimate of a parameter θ, then we say (when we are following the frequentist theory) that

- The true value of the parameter θ is an unknown constant. It is not random.

- An estimate $\widehat{\theta}_n$ of this parameter is a random variable and the correct way to describe its randomness is its sampling distribution.

Frequentist Statistics (cont.)

The frequentist theory has a fundamental problem. It says the correct way to describe the randomness of $\widehat{\theta}_n$ is its sampling distribution, which depends on the parameter θ and perhaps other parameters, the true values of which are unknown.

Thus we seem to be in an infinite regress. Suppose we want to estimate the population mean μ using the sample mean $\widehat{\mu}_n$ as an estimate. When n is large, we know

$$\widehat{\mu}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

but we don't know μ or σ^2. We can estimate both, in fact, we already said $\widehat{\mu}_n$ estimates μ and we can find another estimate $\widehat{\sigma}_n^2$ of σ^2. But now we need to know about the variability of the random vector $(\widehat{\mu}_n, \widehat{\sigma}_n^2)$.

Frequentist Statistics (cont.)

Frequentist statistics is all about how to deal with this infinite regress. Details follow over most of the semester.

Frequentist statistics has nothing whatsoever to do with the frequentist theory of probability, although it is named for it and has been confused with it by famous statisticians.

Frequentist statistics says sampling distributions are the correct measure of uncertainty of estimators. Frequentist probability says probabilities can only be defined with respect to infinite IID sequences.

"Sampling distribution statistics" would be a better name than "frequentist statistics" but you can't change the way people talk.

Bayesian Statistics

Bayesian statistics is name after its originator Thomas Bayes, whose work was published posthumously in 1764.

It makes conditional probability the fundamental tool of inference. It takes probability theory as the correct description of all uncertainty.

If we don't know the true value of a parameter θ, then we are uncertain about it, and the correct description of our knowledge about it or lack thereof is a probability distribution.

What the frequentist calls a statistical model, the Bayesian calls a conditional distribution. The frequentist writes $f_\theta(x)$ for the relevant PDF. The Bayesian writes $f(x \mid \theta)$ for the relevant PDF. Data x and parameter θ are both random to the Bayesian.

Bayes Rule

Before we see data, we have a distribution for θ that reflects our knowledge about it. Say the PDF is $g(\theta)$. This is called the *prior distribution*.

After we see data, we have a joint distribution (marginal times conditional)

$$f(x \mid \theta) g(\theta)$$

and we can find the other conditional

$$f(\theta \mid x) = \frac{f(x \mid \theta) g(\theta)}{\int f(x \mid \theta) g(\theta) \, d\theta}$$

that reflects our knowledge about θ after we see x. This is called the *posterior distribution*.

Bayes Rule (cont.)

The Bayes rule is also called the Bayes theorem, an overly fancy name for some philosophical woof about a straightforward application of the definition of conditional probability.

What is called Bayes rule is the process of finding the other conditional. Given $f(x \mid \theta)$ and $g(\theta)$, find $f(\theta \mid x)$.

Bayesian Statistics

When we are following the Bayesian theory

- The true value of the parameter θ is an unknown constant. Therefore it is random.

- An estimate $\widehat{\theta}_n$ of this parameter is not a random variable after it is seen. The only randomness remaining is in the posterior distribution.

Frequentist versus Bayesian Statistics

The frequentist uses sampling distributions, the Bayesian does not.

The Bayesian uses prior distributions, the frequentist does not.

The frequentist says $\widehat{\theta}_n$ is random but θ is not.

The Bayesian says θ is random but $\widehat{\theta}_n$ is not (after it is seen).

Frequentist versus Bayesian Statistics (cont.)

Bayesian theory is more straightforward than frequentist statistics. All Bayesian inference is application of Bayes rule. Frequentist inference is fragmented. There are dozens of methods.

Bayesian theory is both more difficult and easier than frequentist statistics. Very easy frequentist inferences are moderately hard for the Bayesian. Very difficult or impossible frequentist inferences are moderately hard for the Bayesian.

Frequentist Statistics (cont.)

More on Bayes later. For the next several weeks we do frequentist statistics only.

Until Bayes returns

- Statistics are random variables, their probability distributions are called sampling distributions.

- Parameters are not random variables, they are unknown constants.

Nuisance Parameters

Some parameters are more important than others. Which parameters are more important depends on the context.

The technical jargon for the most important parameter or parameters is *parameter of interest*.

The technical jargon for the other parameter or parameters is *nuisance parameter*.

When we are using $\overline{X}_n$ to estimate μ, we may also need deal with the parameter σ^2, but μ is the parameter of interest and σ^2 is a nuisance parameter.

Estimates (cont.)

Some estimates are better than others. One of the main themes of frequentist statistics is the properties of estimates that make one better than another.

Obviously silly estimates like the constant always equal to 42 are obviously bad. We need theory to help choose among estimates not obviously silly.

Suppose the statistical model under consideration is the family of all probability distributions on $\mathbb{R}$ that are symmetric and have first moments. The parameter of interest is the center of symmetry, which is also the mean and also the median. Thus $\overline{X}_n$ and $\widetilde{X}_n$ are both obvious estimators of this parameter. Which is better? According to what criteria? Under what conditions?

Bias and Unbiasedness

If T is an estimator of a parameter θ, then we say T is *unbiased* if

$$E_\theta(T) = \theta, \qquad \text{for all } \theta \in \Theta$$

where Θ is the parameter space of the statistical model under consideration.

The notation E_θ denotes the expectation operator for the distribution with parameter value θ.

An estimate is unbiased if its expectation is the parameter it estimates.

Bias and Unbiasedness (cont.)

If an estimator is not unbiased, then it is *biased*.

The *bias* of an estimator T of a parameter θ is

$$b(\theta) = E_\theta(T) - \theta$$

Bias and Unbiasedness (cont.)

Many people who have not had a course like this are overly impressed with the concept of unbiasedness.

The concept is very simple. It is the only theoretical concept simple enough to be introduced in elementary courses. It may be the only theoretical concept a theoretically naive person knows.

It is badly named because in ordinary parlance "bias" is bad and "unbiasedness" is good. One expects the same in statistics. But statistical unbiasedness is not a particularly good property.

Bias and Unbiasedness (cont.)

In theoretical statistics unbiasedness is a technical term that allows us to state some theorems concisely. Later we will cover a theorem (the Gauss-Markov theorem) that says the sample mean $\overline{X}_n$ is the best linear unbiased estimator (BLUE) of the population mean μ.

Theoretically naive people, on hearing about this theorem, often think this means that $\overline{X}_n$ is best. Doesn't the theorem say that? The best estimator is $\overline{X}_n$, which is linear and unbiased?

No. The theorem says that among all linear and unbiased estimators, $\overline{X}_n$ is the best. The theorem says nothing about nonlinear or biased estimators. Some of them may be better than $\overline{X}_n$. Probably some are. Otherwise we could state and prove a stronger theorem.

Bias and Unbiasedness (cont.)

Also the Gauss-Markov theorem doesn't say unbiasedness is a good thing.

It *assumes* unbiasedness. It doesn't *conclude* anything about unbiasedness.

We are playing the theoretical game. If we make certain assumptions, we can get certain conclusions. Here we assume unbiasedness and linearity. If we don't assume them, then we don't have a theorem.

Bias and Unbiasedness (cont.)

Sometimes unbiasedness necessitates silliness.

Suppose θ is a parameter known to be nonnegative: the parameter space is $[0, \infty)$.

Suppose $\widehat{\theta}_n$ is an unbiased estimator of θ.

Then we know

$$E_\theta(\widehat{\theta}_n) = \theta, \qquad \theta \geq 0.$$

If $\widehat{\theta}_n$ is non-silly, then it should be nonnegative valued so every estimate is a possible parameter value. But then

$$E_0(\widehat{\theta}_n) = 0$$

implies $\widehat{\theta}_n = 0$ almost surely.

Bias and Unbiasedness (cont.)

Usually the only way to make an estimator constant almost surely is to make it constant period. But then it is silly.

An old friend of mine used to say "I'm ambidextrous, I do equally poorly with both hands".

The "principle" of unbiasedness is doing equally poorly on both sides as a matter of "principle". Stated that way, it is obviously dumb.

If $\widehat{\theta}_n$ is an unbiased estimator of a nonnegative parameter θ, then

$$T = \widehat{\theta}_n I_{[0,\infty)}(\widehat{\theta}_n)$$

is a better estimator. When $\widehat{\theta}_n$ is nonnegative, the two estimators agree. When $\widehat{\theta}_n$ is negative, then T is zero and closer to the true unknown value of θ.

Mean Square Error

The *mean square error* (MSE) of an estimator T of a parameter θ is

$$\text{mse}_\theta(T) = E_\theta\{(T - \theta)^2\}$$

The mean square error formula (5101, deck 2, slides 33-36) says

$$\text{mse}_\theta(T) = \text{var}_\theta(T) + b(\theta)^2$$

where $b(\theta)$ is the bias. In short

$$\text{mean square error} = \text{variance} + \text{bias}^2$$

Mean Square Error (cont.)

MSE is one sensible measure of goodness of an estimator. We will use it a lot.

MSE is another reason why unbiasedness in not necessarily good.

Often there is a bias-variance trade-off. You can make bias small only by increasing variance and vice versa. The only way you can make bias zero is to make variance very large or even infinite. That's not a good trade.

More on the bias-variance trade-off near the end of the semester.

Mean Square Error (cont.)

We already have two estimators of the population variance σ^2 and now we add another

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

$$V_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

$$W_n = \frac{1}{n+1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

S_n^2 is the obvious unbiased estimator. V_n is the variance of the empirical distribution and has nice properties because of that. W_n minimizes MSE if the data are IID normal (homework problem).

Mean Square Error (cont.)

The unbiased estimator is not best if MSE is the criterion (W_n is best).

This is because of bias-variance trade-off. You need some bias to reduce the variance sufficiently to get the smallest MSE.

Mean Square Error (cont.)

Every function of a parameter is a parameter.

It is not true that every function of an unbiased estimator is unbiased (you can't take nonlinear functions out of expectations, 5101, deck 2, slide 5–9).

Mean Square Error (cont.)

S_n^2 is an unbiased estimator of σ^2.

S_n is a biased estimator of σ.

$$\text{var}(S_n) = E(S_n^2) - E(S_n)^2$$

implies

$$E(S_n)^2 = \sigma^2 - \text{var}(S_n)$$

so S_n could only be unbiased if it had variance zero, which would imply it is almost surely constant. But we can't make it constant unless we know the value of the parameter, in which case we don't need to estimate it.

Consistency

A statistic $\widehat{\theta}_n$ is a *consistent* estimator of a parameter θ if

$$\widehat{\theta}_n \xrightarrow{P} \theta, \qquad \text{as } n \to \infty$$

The sample mean is a consistent estimate of the population mean.
The sample median is a consistent estimate of the population median.
The sample variance S_n^2 is a consistent estimate of the population variance σ^2.
The empirical variance V_n is also a consistent estimate of the population variance σ^2.

Consistency (cont.)

Consistency is a very weak property. Really silly estimators like the estimator always equal to 42 are not consistent (unless the true unknown parameter value just happens to be 42).

It is hard to justify using an inconsistent estimator for anything, but just because an estimator is consistent doesn't make it good.

Consistency and Asymptotically Normal

A statistic $\widehat{\theta}_n$ is a *consistent and asymptotically normal* (CAN) estimator of a parameter θ if

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tau^2), \qquad \text{as } n \to \infty$$

for some constant τ^2 that doesn't necessarily have anything to do with the population variance.

The constant τ^2 is called the *asymptotic variance* of the CAN estimator $\widehat{\theta}_n$.

Consistency and Asymptotically Normal (cont.)

The sample mean $\overline{X}_n$ is a CAN estimate of the population mean μ. Its asymptotic variance is the population variance σ^2.

The sample median is a CAN estimate of the population median m. Its asymptotic variance is $1/[4f(m)^2]$, where f is the population PDF.

The sample variance S_n^2 is a CAN estimate of the population variance σ^2. Its asymptotic variance is $\mu_4 - \sigma^4$, where μ_4 is the population fourth central moment.

All already done.

Asymptotic Relative Efficiency

The *asymptotic relative efficiency* (ARE) of two CAN estimators of the same parameter is the ratio of their asymptotic variances.

This is a sensible measure of goodness of an estimator, because if τ^2 is the asymptotic variance of $\widehat{\theta}_n$, this means

$$\widehat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{\tau^2}{n}\right)$$

The actual variance for sample size n is approximately τ^2/n.

Asymptotic Relative Efficiency

So suppose we have two CAN estimators of the same parameter but different asymptotic variances τ_1^2 and τ_2^2 and different sample sizes n_1 and n_2.

Arrange so that the actual variances are approximately equal

$$\frac{\tau_1^2}{n_1} \approx \frac{\tau_2^2}{n_2}$$

Then

$$\frac{\tau_1^2}{\tau_2^2} \approx \frac{n_1}{n_2}$$

The ARE is approximately the ratio of sample sizes need to get the same accuracy, because variance measures the spread-outness of the sampling distribution of an estimator.

Asymptotic Relative Efficiency (cont.)

If cost of data is proportional to sample size (which figures), then ARE is the correct measure of relative cost to get the same accuracy.

When stating an ARE, it is unclear which is better just from the number. Is 2.73 the ARE τ_1^2/τ_2^2 or the ARE τ_2^2/τ_1^2? For clarity state not only the number but also which estimator is better.

The one with the smaller asymptotic variance is better.

Asymptotic Relative Efficiency (cont.)

For a symmetric population distribution having second moments, the sample mean and the sample median are both CAN estimators of the center of symmetry, which is also the mean and the median.

We use ARE to compare them.

The ARE depends on the population distribution.

Asymptotic Relative Efficiency (cont.)

Suppose the population distribution is $\mathcal{N}(\mu, \sigma^2)$.

The sample mean is a CAN estimator of μ. Its asymptotic variance is σ^2.

The sample median is a CAN estimator of μ. Its asymptotic variance is

$$\frac{1}{4f(\mu)^2} = \frac{1}{4 \cdot \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^2} = \frac{\pi\sigma^2}{2}$$

Since $\sigma^2 < \pi\sigma^2/2$, the sample mean is the better estimator. The ARE is $\pi/2$ or $2/\pi$, depending on which you put on top.

Asymptotic Relative Efficiency (cont.)

When the population is normal, the sample mean is a better estimator of μ than the sample median. The ARE is either

$$\frac{\pi}{2} = 1.570796$$

or

$$\frac{2}{\pi} = 0.6366198$$

Asymptotic Relative Efficiency (cont.)

Suppose the population distribution is Laplace(μ, σ^2) using the parametrization on the brand name distributions handout.

The sample mean is a CAN estimator of μ. Its asymptotic variance is σ^2.

The sample median is a CAN estimator of μ. Its asymptotic variance is

$$\frac{1}{4f(\mu)^2} = \frac{1}{4 \cdot \left(\frac{\sqrt{2}}{2\sigma}\right)^2} = \frac{\sigma^2}{2}$$

Since $\sigma^2 > \sigma^2/2$, the sample median is the better estimator. The ARE is 1/2 or 2, depending on which you put on top.

Asymptotic Relative Efficiency (cont.)

When the population is Laplace, also called double exponential, the sample median is a better estimator of μ than the sample mean. The ARE is either 2 or 1/2.

Asymptotic Relative Efficiency (cont.)

We have already seen enough to see that which estimator is better depends on the population distribution.

Sometimes the mean is better (for example, when the population is normal), and sometimes the median is better (for example, when the population is Laplace).

You have to calculate the ARE for each population distribution you want to know about.

Method of Moments

The *method of moments* is a catchphrase for the following style of estimation. We already know

- Every empirical ordinary or central moment of order k is a consistent estimator of the corresponding population moment, assuming that population moments of order k exist (deck 1, slides 90–94).
- Every empirical ordinary or central moment of order k is a CAN estimator of the corresponding population moment, assuming that population moments of order $2k$ exist (deck 1, slides 82–89 and 95–100).

These empirical moments are jointly consistent or jointly CAN, although we didn't prove this.

Method of Moments (cont.)

Apply the continuous mapping theorem to the first and the multivariate delta method to the second, obtaining

- Every continuous function of empirical ordinary or central moments of order k or less is a consistent estimator of the same function of the corresponding population moments, assuming that population moments of order k exist.
- Every differentiable function of empirical ordinary or central moments of order k or less is a CAN estimator of the same function of the corresponding population moments, assuming that population moments of order $2k$ exist.

Method of Moments (cont.)

Thus the method of moments goes as follows. If there are p unknown parameters, choose p moments, evaluate them, this gives p equations giving moments as a function of parameters. Solve these equations for the parameters, so one has p equations giving parameters as a function of moments. Plug in empirical moments for population moments. This gives estimates of the parameters as a function of empirical moments. Derive the asymptotic distribution of the estimators using the delta method.

Method of Moments (cont.)

Our first example is trivial. Suppose X_1, X_2, ... are IID Poi(μ). Find a method of moments estimator for μ. Since there is one parameter, we need one equation, and the obvious one uses the first ordinary moment

$$E_\mu(X) = \mu$$

which says the parameter μ is the identity function of the first ordinary moment μ. Hence the method of moments estimator is

$$\widehat{\mu}_n = \overline{X}_n$$

And, of course, we already know its asymptotic distribution

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\mu}{n}\right)$$

because the population variance is also μ for the Poisson distribution.

Method of Moments (cont.)

Nothing about the method of moments tells us which moments to use. For this Poisson example, we could have used the second central moment

$$\text{var}_\mu(X) = \mu$$

which says the parameter μ is the identity function of the second central moment μ. Hence the method of moments estimator is

$$\widehat{\mu}_n = V_n$$

And, of course, we already know its asymptotic distribution

$$V_n \approx \mathcal{N}\left(\mu, \frac{\mu_4 - \mu^2}{n}\right)$$

where μ_4 is the fourth central moment for the Poisson distribution.

Method of Moments (cont.)

But why would anyone use $\widehat{\mu}_n = V_n$ when $\widehat{\mu}_n = \overline{X}_n$ is much simpler?

We see that there can be many different method of moments estimators for a given problem. But we usually say "the" method of moments estimator, meaning the simplest and most obvious.

Method of Moments (cont.)

Suppose X_1, X_2, ... are IID Exp(λ). Find a method of moments estimator of λ. Since there is one parameter, we need one equation, and the obvious one uses the first ordinary moment

$$E_\lambda(X) = \frac{1}{\lambda}$$

Solving for λ gives

$$\lambda = \frac{1}{E_\lambda(X)}$$

Hence the method of moments estimator is

$$\widehat{\lambda}_n = \frac{1}{\overline{X}_n}$$

Method of Moments (cont.)

We have already worked out its asymptotic distribution (5101, deck 7, slides 56–58)

$$\hat{\lambda}_n \approx \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$$

Method of Moments (cont.)

Suppose X_1, X_2, ... are IID Gam(α, λ). Find method of moments estimators of α and λ. Since there are two parameters, we need two equations, and the obvious ones are

$$E_{\alpha,\lambda}(X) = \frac{\alpha}{\lambda}$$
$$\text{var}_{\alpha,\lambda}(X) = \frac{\alpha}{\lambda^2}$$

It is not always easy to solve simultaneous nonlinear equations, but here

$$\alpha = \frac{E_{\alpha,\lambda}(X)^2}{\text{var}_{\alpha,\lambda}(X)}$$
$$\lambda = \frac{E_{\alpha,\lambda}(X)}{\text{var}_{\alpha,\lambda}(X)}$$

Method of Moments (cont.)

Hence the method of moments estimators for the two-parameter gamma statistical model (two-parameter meaning both parameters unknown) are

$$\widehat{\alpha}_n = \frac{\overline{X}_n^2}{V_n}$$

$$\widehat{\lambda}_n = \frac{\overline{X}_n}{V_n}$$

or

$$\begin{pmatrix} \widehat{\alpha}_n \\ \widehat{\lambda}_n \end{pmatrix} = g(\widehat{\alpha}_n, \widehat{\lambda}_n)$$

where

$$g(u, v) = \begin{pmatrix} u^2/v \\ u/v \end{pmatrix}$$

Method of Moments (cont.)

To find the joint asymptotic normal distribution of these estimators, start with the known asymptotic normal distribution of the empirical moments (deck 1, slide 101)

$$\begin{pmatrix} \overline{X}_n \\ V_n \end{pmatrix} \approx \mathcal{N}\left(\boldsymbol{\mu}, \frac{\mathbf{M}}{n}\right)$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \alpha/\lambda \\ \alpha/\lambda^2 \end{pmatrix}$$

and

$$\mathbf{M} = \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}$$

where we know $\mu_2 = \alpha/\lambda^2$ but need to work out μ_3 and μ_4.

Method of Moments (cont.)

Using the theorem for the gamma distribution and the gamma function recursion formula (both on brand name distributions handout), we know

$$\begin{aligned} E_{\alpha,\lambda}(X^k) &= \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\lambda^k} \\ &= \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{\lambda^k} \end{aligned}$$

(5101, deck 3, slides 140–141).

Method of Moments (cont.)

$$
\begin{aligned}
\mu_3 &= E\left\{\left(X - \frac{\alpha}{\lambda}\right)^3\right\} \\
&= E(X^3) - \frac{3E(X^2)\alpha}{\lambda} + \frac{3E(X)\alpha^2}{\lambda^2} - \frac{\alpha^3}{\lambda^3} \\
&= \frac{\alpha(\alpha+1)(\alpha+2)}{\lambda^3} - \frac{3\alpha^2(\alpha+1)}{\lambda^3} + \frac{3\alpha^3}{\lambda^3} - \frac{\alpha^3}{\lambda^3} \\
&= \frac{2\alpha}{\lambda^3}
\end{aligned}
$$

Method of Moments (cont.)

$$
\begin{aligned}
\mu_4 &= E\left\{\left(X - \frac{\alpha}{\lambda}\right)^4\right\} \\
&= E(X^4) - \frac{4E(X^3)\alpha}{\lambda} + \frac{6E(X^2)\alpha^2}{\lambda^2} - \frac{4E(X)\alpha^3}{\lambda^3} + \frac{\alpha^4}{\lambda^4} \\
&= \frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)}{\lambda^4} - \frac{4\alpha^2(\alpha+1)(\alpha+2)}{\lambda^4} \\
&\quad + \frac{6\alpha^3(\alpha+1)}{\lambda^4} - \frac{4\alpha^4}{\lambda^4} + \frac{\alpha^4}{\lambda^4} \\
&= \frac{3\alpha(\alpha+2)}{\lambda^4}
\end{aligned}
$$

Method of Moments (cont.)

Hence the asymptotic variance matrix is

$$\begin{aligned}\mathbf{M} &= \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix} \\ &= \begin{pmatrix} \alpha/\lambda^2 & 2\alpha/\lambda^3 \\ 2\alpha/\lambda^3 & 2\alpha(\alpha+3)/\lambda^4 \end{pmatrix}\end{aligned}$$

Method of Moments (cont.)

Now we are ready to apply the multivariate delta method (5101, deck 7, slide 100–101) with the change-of-parameter

$$g(u, v) = \begin{pmatrix} u^2/v \\ u/v \end{pmatrix}$$

which has derivative matrix

$$\nabla g(u, v) = \begin{pmatrix} 2u/v & -u^2/v^2 \\ 1/v & -u/v^2 \end{pmatrix}$$

Method of Moments (cont.)

So

$$g\left(\frac{\alpha}{\lambda}, \frac{\alpha}{\lambda^2}\right) = \begin{pmatrix} \alpha \\ \lambda \end{pmatrix}$$

and

$$\nabla g\left(\frac{\alpha}{\lambda}, \frac{\alpha}{\lambda^2}\right) = \begin{pmatrix} 2\lambda & -\lambda^2 \\ \lambda^2/\alpha & -\lambda^3/\alpha \end{pmatrix}$$

Method of Moments (cont.)

Hence (finally!) the asymptotic variance of the method of moments estimators for the two-parameter gamma distribution is (5101, deck 7, slide 100–101)

$$\begin{pmatrix} 2\lambda & -\lambda^2 \\ \lambda^2/\alpha & -\lambda^3/\alpha \end{pmatrix} \begin{pmatrix} \alpha/\lambda^2 & 2\alpha/\lambda^3 \\ 2\alpha/\lambda^3 & 2\alpha(\alpha+3)/\lambda^4 \end{pmatrix} \begin{pmatrix} 2\lambda & \lambda^2/\alpha \\ -\lambda^2 & -\lambda^3/\alpha \end{pmatrix}$$

$$= \begin{pmatrix} 0 & -2\alpha(\alpha+1)/\lambda^2 \\ -1 & -2(\alpha+2)/\lambda \end{pmatrix} \begin{pmatrix} 2\lambda & \lambda^2/\alpha \\ -\lambda^2 & -\lambda^3/\alpha \end{pmatrix}$$

$$= \begin{pmatrix} 2\alpha(\alpha+1) & 2(\alpha+1)\lambda \\ 2(\alpha+1)\lambda & (2\alpha+3)\lambda^2/\alpha \end{pmatrix}$$

Method of Moments (cont.)

In summary,

$$\begin{pmatrix} \widehat{\alpha}_n \\ \widehat{\lambda}_n \end{pmatrix} \approx \mathcal{N}\left[\begin{pmatrix} \alpha \\ \lambda \end{pmatrix}, \frac{1}{n} \begin{pmatrix} 2\alpha(\alpha+1) & 2(\alpha+1)\lambda \\ 2(\alpha+1)\lambda & (2\alpha+3)\lambda^2/\alpha \end{pmatrix} \right]$$

Error Bars

Long before statistics became a subject with its own academic departments, scientists used error bars on plots to show variability of estimators.

Now academic statisticians would say these are approximate, large n, or asymptotic confidence intervals. We will give an official definition of confidence intervals, which was formulated in the 1930's by J. Neyman and E. S. Pearson. But scientists had been using error bars for 200 years before that.

Plug-In

Suppose we have a CAN estimator

$$\widehat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{\tau^2}{n}\right)$$

This says the estimator $\widehat{\theta}_n$ differs from the parameter it estimates θ by about $\tau/\sqrt{n}$ on average. More precisely, the asymptotic normal distribution puts probability

```
Rweb> pnorm(2) - pnorm(-2)
[1] 0.9544997
```

within two standard deviations of the mean, so, when n is large,

$$\Pr\left(|\widehat{\theta}_n - \theta| \le \frac{2\tau}{\sqrt{n}}\right) \approx 0.9544997$$

Plug-In (cont.)

Statisticians are usually fussier. Since

```
Rweb> - qnorm(0.05 / 2)
[1] 1.959964
```

statisticians teach students to say

$$\Pr\left(|\widehat{\theta}_n - \theta| \leq \frac{1.959964\tau}{\sqrt{n}}\right) \approx 0.95$$

but it really doesn't matter. Either statement is only approximate.

Plug-In (cont.)

Now we are faced with the infinite regress problem. We want error bars for θ, which we don't know. But we also don't know any parameters, so we don't know the asymptotic variance τ^2 either. We could estimate that, but then what? An interval for τ^2 that requires its asymptotic variance, which we don't know either? And so forth, ad infinitum?

Fortunately, there is a simple technique, the *plug-in principle*, that saves the day.

Plug-In (cont.)

We have a CAN estimator

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} Y$$

where

$$Y \sim \mathcal{N}(0, \tau^2)$$

Suppose we also have a consistent estimator of the asymptotic variance

$$\hat{\tau}_n^2 \xrightarrow{P} \tau^2$$

Plug-In (cont.)

Combine using Slutsky's theorem

$$\frac{\widehat{\theta}_n - \theta}{\widehat{\tau}_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \frac{Y}{\tau}$$

A linear function of a normal is normal, so Y/τ is normal with parameters

$$\frac{E(Y)}{\tau} = 0$$
$$\frac{\text{var}(Y)}{\tau^2} = 1$$

Thus

$$\frac{\widehat{\theta}_n - \theta}{\widehat{\tau}_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Plug-In (cont.)

The "plug-in" of $\widehat{\tau}_n$ for τ — of a consistent estimator of the asymptotic variance for the asymptotic variance — eliminates the nuisance parameter and eliminates the infinite regress.

Now only the parameter of interest is unknown, and we can make error bars based on

$$\Pr\left(|\widehat{\theta}_n - \theta| \leq \frac{2\widehat{\tau}_n}{\sqrt{n}}\right) \approx 0.9544997$$

or

$$\Pr\left(|\widehat{\theta}_n - \theta| \leq \frac{1.959964\widehat{\tau}_n}{\sqrt{n}}\right) \approx 0.95$$

Error Bars with Plug-In

The usual woof says the error bars have endpoints

$$\widehat{\theta}_n \pm 2 \frac{\widehat{\tau}_n}{\sqrt{n}}$$

or

$$\widehat{\theta}_n \pm 1.96 \frac{\widehat{\tau}_n}{\sqrt{n}}$$

if one is being fussy. Statisticians call the intervals with these endpoints approximate (large n, asymptotic) 95% confidence intervals for θ.

Confidence Intervals

More generally, statisticians allow any probability. If z_α denotes the $1-\alpha$ quantile of the standard normal distribution, then

$$\Pr\left(\left|\frac{\widehat{\theta}_n - \theta}{\widehat{\tau}_n/\sqrt{n}}\right| \leq z_{\alpha/2}\right) \approx 1-\alpha$$

and the interval with endpoints

$$\widehat{\theta}_n \pm z_{\alpha/2}\frac{\widehat{\tau}_n}{\sqrt{n}}$$

is called the asymptotic confidence interval for θ with *coverage probability* $1-\alpha$ or with *confidence level* $100(1-\alpha)\%$.

Confidence Intervals (cont.)

We have now arrived at one of the two key concepts of frequentist statistical inference: confidence intervals and hypothesis tests. To be precise, we have arrived at a special case of confidence intervals. There is more to be said on the subject.

Before getting to that, a few stories.

Confidence Intervals (cont.)

A statistician, Lincoln Moses, had a slogan "use error bars". He said that if he only got that one idea across in intro stats courses, he succeeded.

Then he went to Washington as an official in the Carter administration. When we came back to Stanford, he had a new slogan "use data".

92.3% of all statistics are made up. Especially in politics. Especially in the mainstream media. Or people just tell stories.

Confidence Intervals (cont.)

The worst mistake in statistics. Don't use data, just tell stories. Wrong! Anecdotes are not data.

The second worst mistake in statistics. Don't use error bars. Confuse the sample and the population. Confuse statistics and parameters. Wrong! Sample quantities are only estimates, not truth.

The third worst mistake in statistics. Ignore confidence levels. Wrong! It's called a 95% confidence interval because it misses 5% of the time.

Confidence Intervals (cont.)

A *confidence region* is a random subset of the parameter space — random because it is a function of the data — that has a stated probability of covering the true unknown parameter value. If R is the region, then

$$\Pr_\theta(\theta \in R) = 1 - \alpha, \qquad \text{for all } \theta \in \Theta$$

where Θ is the parameter space and $1 - \alpha$ is the stated coverage probability.

A *confidence interval* is the special case where the parameter is one-dimensional and the set R is always an interval, in which case it can be described by giving two statistics, which are the endpoints of the interval.

Confidence Intervals (cont.)

Those who like to be fussily philosophical take time out for philosophy at this point. If you are a hard-core frequentist, you stress that $\Pr_\theta(\theta \in R)$ does not treat θ as a random variable. That's what Bayesians do, and frequentists are not Bayesians. The random thingummy in $\Pr_\theta(\theta \in R)$ is the region R.

Some fussbudgets will even say it is wrong to say this is the "probability that θ is in R" — instead one must always say the "probability that R contains θ" — even though logically the statements are equivalent. In other words, it is not enough to say what you mean, you must say it in a way that shows your identification with the frequentists.

Confidence Intervals (cont.)

Confidence regions are little used and we will say no more about them. We have an official definition

$$\Pr_\theta(\theta \in R) = 1 - \alpha, \qquad \text{for all } \theta \in \Theta$$

but have no idea how to achieve this in general. Arranging exact equality for all θ ranges from the very difficult to the impossible.

The best one can hope for in most applications is only approximate coverage (replace $=$ with $\approx$) as we did in the large-sample intervals, which are the only ones we have done so far.

Confidence Intervals (cont.)

For discrete data, exact confidence intervals are impossible for a very simple reason. Consider the binomial distribution with sample size n.

There are only $n + 1$ possible data values: 0, 1, ..., n. Hence there are only $n + 1$ possible intervals one can make that are functions of the data. Let R_x denote the confidence interval for data x. The coverage probability is

$$\Pr_p(p \in R) = E_p\{I_R(p)\} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} I_{R_k}(p)$$

Clearly, this piecewise polynomial function of p cannot be constant (equal to $1 - \alpha$ for all p such that $0 < p < 1$).

Confidence Intervals (cont.)

When exact coverage is impossible, one could hope for conservative coverage

$$\Pr_\theta(\theta \in R) \geq 1 - \alpha, \qquad \text{for all } \theta \in \Theta$$

However, this is rarely done. Conservative confidence intervals for the binomial distribution do exist (Clopper-Pearson intervals), but many people argue that they are too conservative, hence more misleading than approximate intervals. So conservative confidence intervals are rarely used for the binomial distribution and never used AFAIK for any other distribution.

We will say no more about conservative intervals. We will either achieve exact coverage or be content with approximate coverage.

Pivotal Quantities

A random variable is called *pivotal* if, firstly, it is a function only of data and the parameter of interest (not a function of nuisance parameters) and, secondly, its distribution does not depend on any parameters.

The most important example is the t pivotal quantity. If the data are IID normal, then

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

(deck 1, slide 76). If μ is the parameter of interest, then the left-hand side contains only data and the parameter of interest and does not contain the nuisance parameter σ^2. The right hand side, the sampling distribution, contains no parameters.

Pivotal Quantities (cont.)

A random variable is called *asymptotically pivotal* if, firstly, it is a function only of data and the parameter of interest (not a function of nuisance parameters) and, secondly, its asymptotic distribution does not depend on any parameters.

The most important examples are those obtained using the plug-in principle

$$\frac{\widehat{\theta}_n - \theta}{\widehat{\tau}_n/\sqrt{n}} \approx \mathcal{N}(0,1)$$

(slide 82). If θ is the parameter of interest, then the left-hand side contains only data and the parameter of interest and does not contain the nuisance parameter τ^2. The right hand side, the asymptotic sampling distribution, contains no parameters.

Pivotal Quantities (cont.)

Sometimes pivotal quantities are called *exact* to distinguish them from asymptotically pivotal quantities. Strictly speaking, the "exact" is redundant. The term "pivotal quantity" means exact pivotal quantity if the adverb "asymptotically" is not attached.

Just like we used asymptotically pivotal quantities to make approximate confidence intervals, we use exact pivotal quantities to make exact confidence intervals.

Exact Confidence Intervals

Let t_α denote the $1-\alpha$ quantile of the $t(n-1)$ distribution. Although this is not indicated by the notation, it does depend on the degrees of freedom.

Since

$$\Pr\left(\left|\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}\right| \leq t_{\alpha/2}\right) = 1-\alpha$$

the interval with endpoints

$$\overline{X}_n \pm t_{\alpha/2}\frac{S_n}{\sqrt{n}}$$

is said to be an exact confidence interval for the population mean with coverage probability $1-\alpha$ or confidence level $100(1-\alpha)\%$.

Exact Confidence Intervals (cont.)

The number $t_{\alpha/2}$ is called a *critical value*.

```
Rweb> qt(1 - 0.05 / 2, 9)
[1] 2.262157
Rweb> qt(1 - 0.05 / 2, 19)
[1] 2.093024
Rweb> qt(1 - 0.05 / 2, 99)
[1] 1.984217
```

are t critical values for 95% confidence and $n = 10$, 20, and 100, respectively. As $n \to \infty$ they converge to

```
Rweb> qnorm(1 - 0.05 / 2)
[1] 1.959964
```

Exact Confidence Intervals (cont.)

```
Rweb> qt(1 - 0.1 / 2, 9)
[1] 1.833113
Rweb> qt(1 - 0.1 / 2, 19)
[1] 1.729133
Rweb> qt(1 - 0.1 / 2, 99)
[1] 1.660391
```

are t critical values for 90% confidence and $n = 10$, 20, and 100, respectively. As $n \to \infty$ they converge to

```
Rweb> qnorm(1 - 0.1 / 2)
[1] 1.644854
```

Exact Confidence Intervals (cont.)

The probability $1 - \alpha$ is called the *coverage probability* in theoretical statistics. In applied statistics, it is always converted to a percentage and called the *confidence level*.

This is the only place percentages are used except in intro stats courses, where some textbook authors think converting probabilities to percentages is helpful rather than confusing — or perhaps those authors think conversion to percentages is a skill students in intro stats courses still need to practice. Converting probabilities to percentages is always confusing IMHO, especially in intro stats courses.

The phrase "95% confidence interval" is so widely used there is no avoiding the percentage there. But don't use percentages anywhere else.

Exact Confidence Intervals (cont.)

A member of my own department actually said to me that these exact t confidence intervals have no valid application because they assume exact normality of the data and no data are ever *exactly* normal. (His research area is probability not statistics, and he doesn't do applied statistics.)

His point is correct as far as it goes. No data are exactly normal. But the t confidence intervals do the right thing when n is large and make a sensible correction when n is small. There is no other method for making such a sensible correction when n is small. Therefore, even if not really "exact" they are still the most sensible thing to do in this situation.

Asymptotic Nonparametric Distribution Free

A method is said to be *asymptotically nonparametric distribution free* if firstly, it works for a nonparametric statistical model (a class of distributions too large to parametrize with a finite set of parameters) and, secondly, it gives asymptotic approximation that does not depend on unknown parameters.

Asymptotic Nonparametric Distribution Free (cont.)

Our main examples so far are the asymptotic confidence intervals based on plug-in. Consider the specific interval

$$\overline{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}}$$

for the population mean. This is asymptotically nonparametric distribution free because it works for a nonparametric statistical model, all distributions having second moments, and the asymptotic distribution

$$\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

does not depend on unknown parameters.

Asymptotic Nonparametric Distribution Free (cont.)

The "exact" t confidence intervals are also asymptotically nonparametric distribution free because

$$t(n-1) \to \mathcal{N}(0,1), \qquad \text{as } n \to \infty$$

hence

$$t_{\alpha/2} \to z_{\alpha/2}, \qquad \text{as } n \to \infty$$

Thus the t confidence intervals are asymptotically equivalent to the plug-in intervals.

Exact Confidence Intervals (cont.)

The fact that t confidence intervals are asymptotically nonparametric distribution free is the first part of our defense of them for practical applications. The second part of our defense is that t critical values are always larger than the normal critical values for the same coverage probability.

When I was a freshman in college, we did some "quantitative analysis" in intro chemistry. We measured the iron content in a little vial of sand we were given. The instructions said to calculate the average and standard deviation of $n = 2$ measurements and report $\overline{X} \pm 2S_n/\sqrt{n}$ as the confidence interval (the chemists didn't teach us about t distributions).

Exact Confidence Intervals (cont.)

If I had known about t distributions back then, I should have used the t critical value

```
Rweb> qt(1 - 0.05 / 2, 1)
[1] 12.70620
```

Clearly, $n = 2$ is not "large" so the interval $\overline{X}_n \pm 2S_n/\sqrt{2}$, which has only asymptotic, large n, justification, is indefensible. The interval $\overline{X}_n \pm 12.7S_n/\sqrt{2}$, being much wider has a much larger coverage probability, which even if not exactly 0.95 will be a lot closer to 0.95 than that of the indefensible interval.

Exact Confidence Intervals (cont.)

Nevertheless, t intervals do not work well unless the population distribution is approximately normal — at least roughly symmetric, unimodal, and light tailed.

Our defense of them is weak. Nonparametric methods (covered later) are better.

Exact Confidence Intervals (cont.)

Assuming the data are IID normal, we can also make confidence intervals for the unknown true population variance σ^2 based on the other pivotal quantity on slide 76

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \text{chi}^2(n-1)$$

Let χ^2_β denote the $1-\beta$ quantile of the $\text{chi}^2(n-1)$ distribution. Although this is not indicated by the notation, it does depend on the degrees of freedom. Then

$$\text{Pr}\left(\chi^2_{1-\alpha+\beta} < \frac{(n-1)S_n^2}{\sigma^2} < \chi^2_\beta\right) = 1-\alpha$$

Exact Confidence Intervals (cont.)

Hence

$$\frac{(n-1)S_n^2}{\chi^2_\beta} < \sigma^2 < \frac{(n-1)S_n^2}{\chi^2_{1-\alpha+\beta}}$$

is an exact $1-\alpha$ confidence interval for σ^2 assuming the data are IID normal. The case $\beta = \alpha/2$ gives *equal tailed* intervals.

If $n = 10$ and $\alpha = 0.05$, the critical values $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are found by

```
Rweb> qchisq(0.05 / 2, 9)
[1] 2.700389
Rweb> qchisq(1 - 0.05 / 2, 9)
[1] 19.02277
```

Exact Confidence Intervals (cont.)

These "exact" confidence intervals for the population variance are not asymptotically nonparametric distribution free. They depend critically on the assumption of normality. From the large degree of freedom approximation for the chi-square distribution

$$\text{chi}^2(n) \approx \mathcal{N}(n, 2n), \qquad \text{when } n \text{ is large}$$

we get

$$\frac{(n-1)S_n^2}{\sigma^2} \approx \mathcal{N}(n, 2n)$$

but this cannot agree with the nonparametric asymptotics

$$S_n^2 \approx \mathcal{N}\left(\sigma^2, \frac{\mu_4 - \sigma^4}{n}\right)$$

because one contains the population forth central moment μ_4 and the other doesn't.

Exact Confidence Intervals (cont.)

Another criticism of intervals based on the chi-square distribution is that there is no particular reason to use equal tailed intervals.

If the reference distribution is symmetric, then equal-tailed make sense. But not otherwise.

Pivotal Quantities Revisited

There is no single best pivotal quantity for a given application. Consider the binomial distribution. We know there can be no exact confidence interval. The obvious method of moments estimator is $\widehat{p}_n = X/n$, where $X \sim \text{Bin}(n, p)$ is the data. The CLT says

$$\sqrt{n}(\widehat{p}_n - p) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, p(1-p)\Big)$$

Since the right-hand side contains parameters, we need to estimate the asymptotic variance, and the obvious estimator is $\widehat{p}_n(1 - \widehat{p}_n)$. Thus the plug-in principle gives

$$\frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1 - \widehat{p}_n)/n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Pivotal Quantities Revisited (cont.)

This gives

$$\widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$$

for the $100(1-\alpha)\%$ asymptotic confidence interval for p.

But we could also use the continuous mapping theorem to obtain another asymptotic pivotal quantity

$$\frac{\widehat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

Pivotal Quantities Revisited (cont.)

The corresponding confidence interval is the set of p that satisfy the inequality

$$\left| \frac{\widehat{p}_n - p}{\sqrt{p(1-p)/n}} \right| \leq z_{\alpha/2}$$

To find that we square both sides and clear the denominator, obtaining

$$n(\widehat{p}_n - p)^2 \leq z^2_{\alpha/2} p(1-p)$$

or

$$n\widehat{p}_n^2 - 2np\widehat{p}_n + np^2 \leq z^2_{\alpha/2} p - z^2_{\alpha/2} p^2$$

Pivotal Quantities Revisited (cont.)

or

$$(n + z_{\alpha/2}^2)p^2 - (2n\widehat{p}_n + z_{\alpha/2}^2)p + n\widehat{p}_n^2 \leq 0$$

The left-hand side is a quadratic function with positive coefficient for the leading term. Hence it goes to ∞ as p goes to $\pm\infty$. Thus the desired interval has endpoints that are the roots of the quadratic equation

$$(n + z_{\alpha/2}^2)p^2 - (2n\widehat{p}_n + z_{\alpha/2}^2)p + n\widehat{p}_n^2 = 0$$

Pivotal Quantities Revisited (cont.)

These endpoints are

$$\frac{2n\hat{p}_n + z^2_{\alpha/2} \pm \sqrt{(2n\hat{p}_n + z^2_{\alpha/2})^2 - 4(n + z^2_{\alpha/2})n\hat{p}_n^2}}{2(n + z^2_{\alpha/2})}$$

$$\frac{2n\hat{p}_n + z^2_{\alpha/2} \pm \sqrt{4n\hat{p}_n z^2_{\alpha/2} + z^4_{\alpha/2} - 4z^2_{\alpha/2} n\hat{p}_n^2}}{2(n + z^2_{\alpha/2})}$$

$$= \frac{\hat{p}_n + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{\left(1 + \frac{z^2_{\alpha/2}}{n}\right)}$$

Pivotal Quantities Revisited (cont.)

Yet another asymptotically pivotal quantity is suggested by the variance stabilizing transformation for the binomial distribution

$$g(p) = \text{asin}(2p - 1)$$

which gives the asymptotically pivotal quantity

$$\sqrt{n}\Big(g(\widehat{p}_n) - g(p)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

(5101, deck 7, slide 67).

Pivotal Quantities Revisited (cont.)

If we define a new parameter

$$\theta = \text{asin}(2p - 1)$$

and its estimator

$$\widehat{\theta}_n = \text{asin}(2\widehat{p}_n - 1)$$

we get

$$\widehat{\theta}_n \pm z_{\alpha/2}\frac{1}{\sqrt{n}}$$

as an asymptotic $100(1-\alpha)\%$ confidence interval for θ.

Pivotal Quantities Revisited (cont.)

If g is an invertible function, and (L, R) is a confidence interval for $g(p)$ for any parameter p, then

$$\left(g^{-1}(L), g^{-1}(R)\right)$$

is the corresponding confidence interval for p if g is increasing, and

$$\left(g^{-1}(R), g^{-1}(L)\right)$$

is the corresponding confidence interval for p if g is decreasing.

Pivotal Quantities Revisited (cont.)

In this case

$$\theta = \text{asin}(2p - 1)$$

solved for p is

$$p = \frac{1 + \sin(\theta)}{2}$$

so

$$g^{-1}(\theta) = \frac{1 + \sin(\theta)}{2}$$

so

$$\frac{1 + \sin(\hat{\theta}_n \pm z_{\alpha/2}/\sqrt{n})}{2}$$

are the endpoints of an asymptotic $100(1 - \alpha)\%$ confidence interval for p.

Pivotal Quantities Revisited (cont.)

So now we have three equally good asymptotic confidence intervals for the parameter of the binomial distribution — equally good as far as asymptotics can tell us.

Which should we use? No theory tells us how these work in practice, when n hasn't gone to infinity. We can simulate various cases and see how each of the intervals works in each simulation. People have done that and concluded that the interval

$$\frac{\widehat{p}_n + \dfrac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{\dfrac{\widehat{p}_n(1 - \widehat{p}_n)}{n} + \dfrac{z_{\alpha/2}^2}{4n^2}}}{\left(1 + \dfrac{z_{\alpha/2}^2}{n}\right)}$$

is the best of the three.

Pivotal Quantities Revisited (cont.)

Let's try them out. The first is done by

```
Rweb> n <- 100
Rweb> x <- 4
Rweb> phat <- x / n
Rweb> phat + c(-1, 1) * qnorm(0.975) * sqrt(phat * (1 - phat) / n)
[1] 0.001592707 0.078407293
```

Pivotal Quantities Revisited (cont.)

With `x` and `n` defined as before, the second is done by

```
Rweb> prop.test(x, n, correct = FALSE)

        1-sample proportions test without continuity correction

data:  x out of n, null probability 0.5
X-squared = 84.64, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.01566330 0.09837071
sample estimates:
   p
0.04
```

Pivotal Quantities Revisited (cont.)

With `x` and `n` defined as before, the third is done by

```
Rweb> thetahat <- asin(2 * phat - 1)
Rweb> (1 + sin(thetahat + c(-1, 1) * qnorm(0.975) / sqrt(n))) / 2
[1] 0.01064524 0.08696897
```

Pivotal Quantities Revisited (cont.)

The three asymptotic 95% confidence intervals

$$(0.0016, 0.0784)$$
$$(0.0157, 0.0984)$$
$$(0.0106, 0.0870)$$

do not differ by much numerically, but their actual achieved coverage probabilities may differ a lot (only a simulation study can tell that).

There is no one right way to do a confidence interval.

Two-Sample Confidence Intervals

Often we don't want a confidence interval for a parameter but for the difference of parameter values for two samples.

These come in two forms: paired comparisons and two independent samples.

Paired Comparisons

In paired comparisons, the data come in pairs. The pairs are assumed IID, but the two components of each pair are not assumed independent. In fact paired comparison procedures work best if the pairs are highly positively correlated. This is often arranged by making the components of each pair two measurements on the same individual (before and after, left and right, treatment and control, etc.)

All paired comparison procedures use a simple trick. If the pairs are (X_i, Y_i), $i = 1, \ldots, n$, then form the differences

$$Z_i = X_i - Y_i$$

which are IID. Analyze the Z_i. This is a one-sample procedure just like the one-sample procedures we have already done.

Two Independent Samples

If X_1, ..., X_m are IID and Y_1, ..., Y_n are IID and both samples are independent of each other. Then

$$E(\overline{X}_m - \overline{Y}_n) = \mu_X - \mu_Y$$
$$\text{var}(\overline{X}_m - \overline{Y}_n) = \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}$$

where μ_X and μ_Y denote the means of the X_i and Y_i, respectively, and similarly for the variances.

From this it seems reasonable that

$$\overline{X}_m - \overline{Y}_n \approx \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)$$

when m and n are large. This is true, although with two different sample sizes m and n, we don't have the tools to prove it.

Two Independent Samples

From this result the asymptotically pivotal quantity resulting from plug-in is

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

where $S^2_{X,m}$ and $S^2_{Y,n}$ are sample variances for the two samples. This gives

$$\overline{X}_m - \overline{Y}_n \pm z_{\alpha/2}\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}$$

as an asymptotic $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$.

Two Independent Samples (cont.)

Things become considerably more complicated if we want an exact confidence interval in the case of two independent samples. We must assume each of the two independent samples are IID normal. Then we know

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0, 1)$$

$$\frac{(m-1)S_{X,m}^2}{\sigma_X^2} + \frac{(n-1)S_{Y,n}^2}{\sigma_Y^2} \sim \text{chi}^2(m + n - 2)$$

and these pivotal quantities are independent.

Two Independent Samples (cont.)

But when we try to make a t random variable out of them

$$T = \frac{\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}}{\sqrt{\frac{\frac{(m-1)S_{X,m}^2}{\sigma_X^2} + \frac{(n-1)S_{Y,n}^2}{\sigma_Y^2}}{m+n-2}}} \sim t(m+n-2)$$

the nuisance parameters don't cancel unless we assume $\sigma_X = \sigma_Y$.

Two Independent Samples (cont.)

So assume $\sigma_X = \sigma_Y$. Then

$$T = \frac{\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{m} + \frac{1}{n}}}}{\sqrt{\frac{(m-1)S^2_{X,m} + (n-1)S^2_{Y,n}}{m+n-2}}} \sim t(m+n-2)$$

is a pivotal quantity if $\mu_X - \mu_Y$ is the parameter of interest. And we can use it to make an exact confidence interval for this parameter.

Two Independent Samples (cont.)

To clean up the math, most intro stats books introduce

$$S^2_{\text{pooled}} = \frac{(m-1)S^2_{X,m} + (n-1)S^2_{Y,n}}{m+n-2}$$

Then

$$T = \frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{S_{\text{pooled}}\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

and

$$\overline{X}_m - \overline{Y}_n \pm t_{\alpha/2} S_{\text{pooled}} \sqrt{\frac{1}{m}+\frac{1}{n}}$$

is an exact $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$, where the t critical value uses $m+n-2$ degrees of freedom.

Two Independent Samples (cont.)

Criticism: not only does this "exact" interval suffer all the problems of the one-sample "exact" procedure — each independent sample must be exactly IID normal for this procedure to be exact — it also suffers from the additional assumption that the two population variances are exactly equal.

The assumption $\sigma_X = \sigma_Y$ is unverifiable in small samples and unnecessary in large samples, because the large-sample procedure does not need it.

Two Independent Samples (cont.)

The "exact" two-sample procedure is not asymptotically equivalent to the large-sample procedure because they use different estimates of the variance of $\overline{X}_m - \overline{Y}_n$. The "exact" procedure uses $S^2_{\text{pooled}}(1/m + 1/n)$, and the large-sample procedure uses $S^2_X/m + S^2_Y/n$.

Hence the "exact" procedure is not asymptotically nonparametric distribution free for the family of all distributions with second moments. The exact confidence interval for the difference of means is only asymptotically nonparametric distribution free under the additional assumption $\sigma_X = \sigma_Y$. The exact confidence interval for the variance is only asymptotically nonparametric distribution free under the additional assumption $\mu_4 = 3\sigma^4$. But these "assumptions" are never valid in real applications.

Two Independent Samples (cont.)

Recognizing the problems with the "exact" procedure, textbooks have recently — only in the last 20 years — stopped recommending it. Instead, an approximate but much better procedure invented by Welch about 50 years ago is now recommended.

The idea is to use the asymptotic pivotal quantity

$$T = \frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}} \approx \mathcal{N}(0, 1)$$

but get a better approximation to its sampling distribution.

We still assume each sample is IID normal. Then

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0, 1) \qquad (*)$$

Two Independent Samples (cont.)

And we see that T is the quotient of $(*)$ and

$$\sqrt{\frac{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}} \qquad (**)$$

And we know that $(*)$ and $(**)$ are independent random variables.

> Assumption: $(**)$ is the square root of a chi-square random variable divided by its degrees of freedom.

If this assumption held, then T would have a t distribution. But it does not hold. We assume it holds *approximately*. Then what t distribution do we get for the approximation?

Two Independent Samples (cont.)

Welch decided to match moments. If $Y \sim \text{chi}^2(\nu)$, then

$$E(Y) = \nu$$
$$\text{var}(Y) = 2\nu$$

Hence

$$E\left(\frac{Y}{\nu}\right) = 1$$
$$\text{var}\left(\frac{Y}{\nu}\right) = \frac{2}{\nu}$$

Lemma

From

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \text{chi}^2(n-1)$$

we derive

$$\text{var}\left(\frac{(n-1)S_n^2}{\sigma^2}\right) = \frac{(n-1)^2\,\text{var}(S_n^2)}{\sigma^4} = 2(n-1)$$

hence

$$\text{var}(S_n^2) = \frac{2\sigma^4}{n-1}$$

Two Independent Samples (cont.)

So what are the corresponding moments of the square of (**)?

$$E\left(\frac{\frac{S^2_{X,m}}{m}+\frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m}+\frac{\sigma^2_Y}{n}}\right)=1$$

$$\text{var}\left(\frac{\frac{S^2_{X,m}}{m}+\frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m}+\frac{\sigma^2_Y}{n}}\right)=\frac{\frac{\text{var}(S^2_{X,m})}{m^2}+\frac{\text{var}(S^2_{Y,n})}{n^2}}{\left(\frac{\sigma^2_X}{m}+\frac{\sigma^2_Y}{n}\right)^2}$$

$$=\frac{\frac{2\sigma^4_X}{m^2(m-1)}+\frac{2\sigma^4_Y}{n^2(n-1)}}{\left(\frac{\sigma^2_X}{m}+\frac{\sigma^2_Y}{n}\right)^2}$$

Two Independent Samples (cont.)

Hence if $(**)$ were the square root of a chi-square divided by its degrees of freedom ν and consequently T would be $t(\nu)$ distributed, then ν would satisfy

$$\frac{2}{\nu} = \frac{\frac{2\sigma_X^4}{m^2(m-1)} + \frac{2\sigma_Y^4}{n^2(n-1)}}{\left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)^2}$$

so

$$\nu = \frac{\left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)^2}{\frac{1}{m-1} \cdot \left(\frac{\sigma_X^2}{m}\right)^2 + \frac{1}{n-1} \cdot \left(\frac{\sigma_Y^2}{n}\right)^2}$$

Two Independent Samples (cont.)

Since ν is a function of the nuisance parameters, we do not know its value. Thus we estimate it

$$\widehat{\nu} = \frac{\left(\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}\right)^2}{\frac{1}{m-1} \cdot \left(\frac{S^2_{X,m}}{m}\right)^2 + \frac{1}{n-1} \cdot \left(\frac{S^2_{Y,n}}{n}\right)^2}$$

This gives Welch's approximate $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$

$$\overline{X}_m - \overline{Y}_n \pm t_{\alpha/2} \sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}$$

where the t critical value uses the t distribution with $\widehat{\nu}$ degrees of freedom.

Two Independent Samples (cont.)

R has a function that makes t confidence intervals

```
Rweb> x <- c(7.7, 8.5, 8.9, 9.7, 10.9, 11.4, 12.6)
Rweb> t.test(x)

        One Sample t-test

data:  x
t = 15.0611, df = 6, p-value = 5.4e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  8.33945 11.57484
sample estimates:
mean of x
 9.957143
```

Two Independent Samples (cont.)

```
Rweb> x <- c(7.7, 8.5, 8.9, 9.7, 10.9, 11.4, 12.6)
Rweb> y <- c(12.1, 13.0, 16.5, 17.9, 21.9)
Rweb> t.test(x, y, var.equal = TRUE)

        Two Sample t-test

data:  x and y
t = -3.7978, df = 10, p-value = 0.003499
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.032446  -2.613268
sample estimates:
mean of x mean of y
 9.957143 16.280000
```

Two Independent Samples (cont.)

```
Rweb> x <- c(7.7, 8.5, 8.9, 9.7, 10.9, 11.4, 12.6)
Rweb> y <- c(12.1, 13.0, 16.5, 17.9, 21.9)
Rweb> t.test(x, y)

        Welch Two Sample t-test

data:  x and y
t = -3.3504, df = 5.13, p-value = 0.01954
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.137153  -1.508561
sample estimates:
mean of x mean of y
 9.957143 16.280000
```

Two Independent Samples (cont.)

Compare. The "exact" 95% confidence interval for $\mu_X - \mu_Y$ that assumes exact normality of both populations and equality of population variances $\sigma_X^2 = \sigma_Y^2$

$$(-10.03, -2.61)$$

and the approximate 95% confidence interval for $\mu_X - \mu_Y$ that also assumes exact normality of both populations

$$(-11.14, -1.51)$$

Not a huge difference, but the later is more defensible because it does not assume $\sigma_X^2 = \sigma_Y^2$.

Hypothesis Tests

Often confidence intervals do not do exactly what is wanted. There are two related reasons for this.

Sometimes the size of an effect is not interesting, only the existence of the effect. In the U. S. A. a drug may be approved for marketing if it is safe and effective. The size of the treatment effect is irrelevant.

Sometimes the size of the effect depends on the details of the particular experiment and would not generalize to other situations. A phenomenon is hypothesized. An experiment is designed to study it. If the experiment shows that the phenomenon exists, then that generalizes to other situations, but the effect size seen does not generalize.

Hypothesis Tests (cont.)

To relate these considerations to statistics, we need to turn these statements about existence of effects and phenomena into statements about a statistical model.

This is often a hard step for scientists, even ones who know better. Scientists want to talk about reality not about statistical models. But statistics only applies to statistical models.

A "statement about a statistical model" is called a *statistical hypothesis*, and formally statistical tests are called *tests of statistical hypotheses* (the plural of hypothesis is hypotheses, the last syllable pronounced like "seas").

Hypothesis Tests (cont.)

A *statistical hypothesis* asserts that the true unknown distribution lies in a submodel of the statistical model under consideration.

If the model under consideration has parameter space Θ, then a statistical hypothesis asserts that the true unknown parameter value lies in a subset Θ_i of Θ.

A hypothesis test considers two hypotheses, conventionally called the *null hypothesis* and the *alternative hypothesis*. As statements they are conventionally denoted H_0 and H_1. As subsets of the parameter space, they are conventionally denoted Θ_0 and Θ_1.

Hypothesis Tests (cont.)

Time out from theory for a concrete example.

We have two independent samples, X_1, ..., X_m and Y_1, ..., Y_n, which are control and treatment, respectively, in a medical experiment. Suppose we are willing to assume that both samples are IID normal. The question of scientific interest is whether the treatment has an effect. We turn this into a question about statistical models: whether μ_X is less than μ_Y. Thus our statistical hypotheses can be

$$H_0 : \mu_X \geq \mu_Y$$
$$H_1 : \mu_X < \mu_Y$$

Hypothesis Tests (cont.)

It turns out, although this is not obvious, that

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X < \mu_Y$$

determine the same hypothesis test and do so more simply. Thus we start with these.

We base the test on Welch's approximate pivotal quantity

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}} \approx t(\widehat{\nu})$$

where $\widehat{\nu}$ is given on slide 142.

Hypothesis Tests (cont.)

Under H_0 this approximate pivotal quantity does not contain parameters, hence is a statistic

$$T = \frac{\overline{X}_m - \overline{Y}_n}{\sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}}$$

which is called the *test statistic*.

Under H_0 the test statistic has sampling distribution centered at zero. Under H_1 the test statistic has sampling distribution centered at some negative number. Thus large negative values of T are evidence in favor of H_1.

Under H_0 we know approximately the distribution of T. Under H_1 we do not, because it depends on μ_X, μ_Y, σ_X^2 and σ_Y^2. Thus we base the probability calculation we do on H_0.

Hypothesis Tests (cont.)

The *P-value* of the test is

$$\Pr(T \le t)$$

where t is the observed value of the test statistic, considered to be nonrandom, and T is the test statistic considered as a random variable, and where the probability is calculated under H_0.

Under H_1, the observed value t is likely to be large and negative, hence far out in the tail of the distribution of T under H_0. Hence the P-value should be small when H_1 is true, but should be large (near 1/2) when H_0 is true.

Thus small P-values are evidence in favor of H_1 and large P-values are evidence in favor of H_0.

Hypothesis Tests (cont.)

```
Rweb> x <- c(7.7, 8.5, 8.9, 9.7, 10.9, 11.4, 12.6)
Rweb> y <- c(12.1, 13.0, 16.5, 17.9, 21.9)
Rweb> t.test(x, y, alternative = "less")

        Welch Two Sample t-test

data:  x and y
t = -3.3504, df = 5.13, p-value = 0.009769
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -2.541343
sample estimates:
mean of x mean of y
 9.957143 16.280000
```

Hypothesis Tests (cont.)

From the on-line help obtained by the command `help(t.test)` in the "Arguments" section

```
alternative: a character string specifying the alternative
          hypothesis, must be one of '"two.sided"' (default),
          '"greater"' or '"less"'.  You can specify just the
          initial letter.
```

and from the "Details" section

```
     'alternative = "greater"' is the alternative that
     'x' has a larger mean than 'y'.
```

Hence `alternative = "greater"` specifies $H_1 : \mu_X > \mu_Y$ so we want `alternative = "less"`, which specifies $H_1 : \mu_X < \mu_Y$.

Hypothesis Tests (cont.)

The statistical analysis ends with the report of the P-value, usually tersely

> Welch two-sample t-test, $t = -3.35$, $P = 0.0098$

The scientific analysis then resumes. Since the P-value is small, this is evidence in favor of H_1. The scientific interpretation of this is that the treatment effect does exist.

Hypothesis Tests (cont.)

Because the P-value could be smaller, the evidence could be stronger, hence the evidence is not absolutely conclusive.

But we can say — assuming the data are IID normal in each sample and the samples are independent — that either H_1 is true or a rather unusual event has occurred, since the event $T \leq t$ occurs with probability 0.0098 when H_0 is true.

Hypothesis Tests (cont.)

Had the evidence come out differently, say $P = 0.098$, this would be much weaker evidence in favor of H_1.

We can still say that either H_1 is true or a somewhat unusual event has occurred, since the event $T \leq t$ occurs with probability 0.098 when H_0 is true, but roughly one time out of ten is not very unusual.

Hypothesis Tests (cont.)

You don't have to talk to many people about statistics before noticing that the number 0.05 is held in quasi-religious awe by many. If $P \leq 0.05$ the result is declared to be "statistically significant" and treated with great respect.

The number 0.05 is clearly arbitrary, considered a round number because people have five fingers. Computers, which count in binary, would consider 1/16 or 1/32 round numbers but would not consider $0.05 = 1/20$ a round number.

Anyone who considers $P = 0.051$ and $P = 0.049$ radically different understands neither science nor statistics.

Hypothesis Tests (cont.)

Yet many scientists, including journal editors and referees, do seem to act as if they consider $P = 0.051$ and $P = 0.049$ radically different.

This is partly the result of bad statistics teaching, and partly the very human wish for a definite conclusion — asking statistics for what it cannot deliver.

People want a sharp dividing line. Either the experiment demonstrates the effect or it doesn't. But statistics only deals in probabilities. The smaller the P-value the stronger the evidence in favor of the alternative hypothesis, but there is no P-value that absolutely establishes the truth of H_1.

Hypothesis Tests (cont.)

Some journals have explicitly stated that papers must say $P < 0.05$ to be publishable. It is widely believed that most journals do likewise.

This leads to the following quip

> Statistics is the branch of applied mathematics that allows one to do twenty bad experiments and get one paper in *Nature*.

Hypothesis Tests (cont.)

The more you think about this joke, the more disturbing it is.

Many small studies are done. Some have $P < 0.05$ by chance alone. If only those papers are published, then *all published papers* about one issue point in the same direction. But this is entirely due to the publication process and has nothing to do with scientific reality.

Refusal to publish papers saying $P > 0.05$ is refusal to publish contrary evidence. Nothing could be more unscientific.

Hypothesis Tests (cont.)

This is only beginning to be recognized. Habits are hard to change, even among scientists.

The study of published literature, the attempt to synthesize the results of many studies, is called *meta-analysis*. In meta-analysis, the tendency to publish only papers saying $P < 0.05$ is called the *file drawer problem*. If all the studies with $P > 0.05$ remain in file drawers rather than being published, then the meta-analyst must treat them as missing data.

Many studies with P-values only slightly below 0.05 are actually fairly strong contrary evidence (against H_1), because P-values should follow a continuous distribution so unpublished P-values slightly above 0.05 must also have been common.

Hypothesis Tests (cont.)

In some areas, every experiment started is recorded in a registry. Thus meta-analysts know the total number of experiments and can conclude that the unpublished ones did not favor the alternative hypothesis.

But this is still uncommon.

Hypothesis Tests (cont.)

We now return to the formal theory.

In general Θ_0 and Θ_1 can be any two disjoint subsets of the parameter space.

When Θ_0 is a singleton set (contains exactly one point), the null hypothesis is said to be *simple*.

If the test is based on a test statistic, the alternative hypothesis plays no role other than motivating the choice of test statistic.

Hypothesis Tests (cont.)

In formal theory we usually assume that large values of the test statistic favor H_1. Then for a simple null hypothesis

$$H_0 : \theta = \theta_0$$

the P-value is

$$\Pr_{\theta_0}(T \geq t)$$

In practice, we often allow the test statistic to not have this form. Then the theory needs to be adjusted correspondingly.

One-Tailed and Two-Tailed Tests

In tests where the hypotheses involve a single parameter θ, and the distribution of the test statistic T is symmetric about zero under H_0, we distinguish three kinds of tests.

Upper-Tailed Tests

The hypotheses are

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$

and the P-value is

$$\Pr_{\theta_0}(T \geq t)$$

One-Tailed and Two-Tailed Tests (cont.)

Lower-Tailed Tests

The hypotheses are

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta < \theta_0$$

and the P-value is

$$\Pr_{\theta_0}(T \leq t)$$

One-Tailed and Two-Tailed Tests (cont.)

Two-Tailed Tests

The hypotheses are

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

and the P-value is

$$\begin{aligned} \Pr_{\theta_0}(|T| \geq |t|) &= \Pr_{\theta_0}(T \leq -|t| \text{ or } T \geq |t|) \\ &= 2 \Pr_{\theta_0}(T \leq -|t|) \\ &= 2 \Pr_{\theta_0}(T \geq |t|) \end{aligned}$$

One-Tailed and Two-Tailed Tests (cont.)

(Still assuming the distribution of T is symmetric about zero under H_0) the P-values for the one-tailed tests always add to one, so one is less than $1/2$ and the other greater than $1/2$. At most one can give a statistically significant result.

The P-value for the two-tailed test is always twice the P-value for the one-tailed that is less than $1/2$.

Hence the P-value for two-tailed test usually looks less significant than the P-value for a one-tailed test.

The Dogma of Hypothesis Tests

Do only one test per data set.

Not just report only one test, *do* only one test.

Moreover the test to be done is chosen before the data are observed.

This dogma is often violated but unrestricted multiple testing without correction makes statistics like playing tennis without a net — an entirely meaningless exercise. More on this later.

One-Tailed and Two-Tailed Tests (cont.)

According to the dogma, a one-tailed test is valid if you

- choose which tail before the data are collected, and,

- if $P > 1/2$, then that is the end. No more statistics is done. The data are thrown in the trash. Nothing is published.

One-Tailed and Two-Tailed Tests (cont.)

Whether a one-tailed test is valid is a scientific question not a statistical one.

A one-tailed test is valid if readers can believe the analysis was done according to the dogma. If readers suspect that the experiment would have been published regardless of which alternative the evidence had favored, then that is tantamount to a suspicion that the one-tailed test is invalid.

One-Tailed and Two-Tailed Tests (cont.)

Sometimes it is clear that a one-tailed test is valid. If the alternatives are treatment better than control and treatment worse than control, then there is no question that the latter would not have lead to publication.

Sometimes it is clear that a one-tailed test is not valid. If the alternatives are men better than women and women better than men, then there is a fair question that the either would have lead to publication.

Sometimes it is unclear. Readers have to make up their own minds.

When a reader thinks a one-tailed test invalid, the reader can convert it to a two-tailed test by doubling the P-value.

One-Tailed and Two-Tailed Tests (cont.)

There is a pernicious connection between worship of the number 0.05 and one-tailed test. A scientist straining to obtain $P < 0.05$ will sometimes switch from two-tailed to one-tailed — thereby cutting the P-value in half — to obtain $P < 0.05$.

This is bogus. I call it "honest cheating" because there is no fraud (if it is clearly stated that a one-tailed test was done). The bogosity is clear to expert readers, who mentally double the P-value. Naive readers are fooled.

Decision Theory

The theory of statistical decisions is a large subject, and we will only look at a little part. When applied to hypothesis tests, it gives a different view.

The point of a hypothesis test is to decide in favor of H_0 or H_1. The result is one of two decisions, conventionally called

- accept H_0 or reject H_1 (both mean the same)
- reject H_0 or accept H_1 (both mean the same)

In the decision-theoretic mode, the result of a test is just reported in these terms. No P-value is reported, hence no indication of the strength of evidence.

Decision Theory (cont.)

If no P-value is reported, then how is the test done?

A *level of significance* α is chosen.

- If $P < \alpha$, then the test decides "reject H_0".
- If $P \geq \alpha$, then the test decides "accept H_0".

It is clear that the decision theoretic view simply provides less information to readers. Instead of giving the actual P-value, it is only reported whether the P-value is above or below α.

Decision Theory (cont.)

Ideally, the significance level α should be chosen carefully and reflect the costs and probabilities of false positive and false negative decisions.

In practice $\alpha = 0.05$ is usually thoughtlessly chosen.

Since the decision-theoretic mode provides less information and isn't usually done properly anyway, many recent textbooks say it should not be used: always report the P-value, never report only a decision.

Decision Theory (cont.)

So why hasn't the decision-theoretic mode gone away?

The decision-theoretic mode makes for much simpler theory. P-values can be hard to define in complicated situations when there is no obvious choice of test statistic.

If one knows how to do a test for any α between zero and one, then for any given data the test will accept H_0 when α is small and reject H_0 when α is large, and the P-value can be defined as the number that separates these two regions. The P-value is the infimum of α for which the test rejects H_0 or the supremum of α for which the test accepts H_0.

Decision Theory (cont.)

There is also a bad reason why textbooks still teach the decision-theoretic mode. It was there first. P-values came later. Textbooks are often decades behind current trends. Many teachers are decades out of date. Many referees and editors of journals are decades out of date on many things. They are, of course, expert in their areas, but they may not be experts in statistics.

One-Tailed and Two-Tailed Tests (cont.)

When there is a compound null hypothesis, it is not clear how to define the P-value. One definition is

$$\sup_{\theta \in \Theta_0} \Pr_\theta(T \geq t)$$

Now the P-value is no longer a probability.

The corresponding decision-theoretic view is

$$\sup_{\theta \in \Theta_0} \Pr_\theta(\text{reject } H_0) \leq \alpha$$

however the decision "reject H_0" is determined.

One-Tailed and Two-Tailed Tests (cont.)

Consider a one-tailed test based on the exact or asymptotic pivotal quantity

$$\frac{\widehat{\theta}_n - \theta}{\widehat{\tau}_n / \sqrt{n}}$$

and compound null hypothesis

$$H_0 \colon \theta \leq \theta_0$$
$$H_1 \colon \theta > \theta_0$$

We claim the test having the test statistic

$$T = \frac{\widehat{\theta}_n - \theta_0}{\widehat{\tau}_n / \sqrt{n}}$$

and P-value $\Pr_{\theta_0}(T \geq t)$ is valid.

One-Tailed and Two-Tailed Tests (cont.)

We must show that

$$\Pr_\theta(T \geq t) \leq \Pr_{\theta_0}(T \geq t), \qquad \theta < \theta_0$$

If the true unknown parameter value is θ, then

$$\frac{\widehat{\theta}_n - \theta}{\widehat{\tau}_n/\sqrt{n}} = \frac{\widehat{\theta}_n - \theta_0}{\widehat{\tau}_n/\sqrt{n}} + \frac{\theta_0 - \theta}{\widehat{\tau}_n/\sqrt{n}} = T + \frac{\theta_0 - \theta}{\widehat{\tau}_n/\sqrt{n}}$$

has the same distribution as T does when the true unknown parameter value is θ_0. Hence

$$\Pr_\theta\left(T + \frac{\theta_0 - \theta}{\widehat{\tau}_n/\sqrt{n}} \geq s\right) = \Pr_{\theta_0}(T \geq s)$$

or

$$\Pr_\theta(T \geq t) = \Pr_{\theta_0}\left(T \geq t + \frac{\theta_0 - \theta}{\widehat{\tau}_n/\sqrt{n}}\right)$$

One-Tailed and Two-Tailed Tests (cont.)

We assume $\widehat{\tau}_n > 0$, which makes sense since it is an estimate of standard deviation, then if H_0 is true (so $\theta \leq \theta_0$)

$$\frac{\theta_0 - \theta}{\widehat{\tau}_n / \sqrt{n}} \geq 0$$

and

$$\Pr_\theta(T \geq t) = \Pr_{\theta_0}\left(T \geq t + \frac{\theta_0 - \theta}{\widehat{\tau}_n / \sqrt{n}}\right) \leq \Pr_{\theta_0}(T \geq t)$$

One-Tailed and Two-Tailed Tests (cont.)

In conclusion: the test with P-value

$$\Pr_{\theta_0}(T \geq t)$$

is valid for either

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \leq \theta_0$$
$$H_1 : \theta > \theta_0$$

whether the null hypothesis is an equality or inequality is irrelevant. And similarly for the other one-tailed test.

Power

Everything we have said about hypothesis tests so far is only about validity. Is the test defensible?

A different issue is power, which is, roughly, how probable is it that the test will do what is wanted.

More precisely, the *power* of a test is the probability that it will accept H_1 when H_1 is true. Since H_1 is always a composite hypothesis, the power is always a function of the true unknown parameter value θ. It also depends on the sample size.

Since the issue is about accepting H_1, this inherently takes the decision-theoretic viewpoint.

Power (cont.)

The power of a hypothesis is useful in planning an experiment or getting funding for an experiment. Most grant proposals include power calculations. What would be the point of funding an experiment that probably won't detect anything anyway because the planned sample size is too small?

In order to do a power calculation we need to specify certain values for the parameters and the sample size.

Hence a power calculation is hypothetical. It assumes certain values of the parameters, which must be made up.

Power (cont.)

Power calculations are simplest when the reference distribution (the distribution of the test statistic under H_0, either exact or approximate) is normal. We start there, considering again the situation on slides 182–184. The asymptotically pivotal quantity is

$$\frac{\widehat{\theta}_n - \theta}{\widehat{\tau}_n/\sqrt{n}} \approx \mathcal{N}(0,1)$$

the test statistic is

$$T = \frac{\widehat{\theta}_n - \theta_0}{\widehat{\tau}_n/\sqrt{n}}$$

and we found out

$$\mathsf{Pr}_\theta(T \geq t) = \mathsf{Pr}_{\theta_0}\left(T \geq t + \frac{\theta_0 - \theta}{\widehat{\tau}_n/\sqrt{n}}\right)$$

Power (cont.)

In

$$\mathsf{Pr}_\theta(T \geq t) = \mathsf{Pr}_{\theta_0}\left(T \geq t + \frac{\theta_0 - \theta}{\widehat{\tau}_n/\sqrt{n}}\right)$$

the plug-in does not help, since we do not know the sampling distribution of $\widehat{\tau}_n$ (we only know that it is a consistent estimator of the nuisance parameter τ). So we write

$$\mathsf{Pr}_\theta(T \geq t) \approx \mathsf{Pr}_{\theta_0}\left(T \geq t + \frac{\theta_0 - \theta}{\tau/\sqrt{n}}\right)$$

undoing the plug-in.

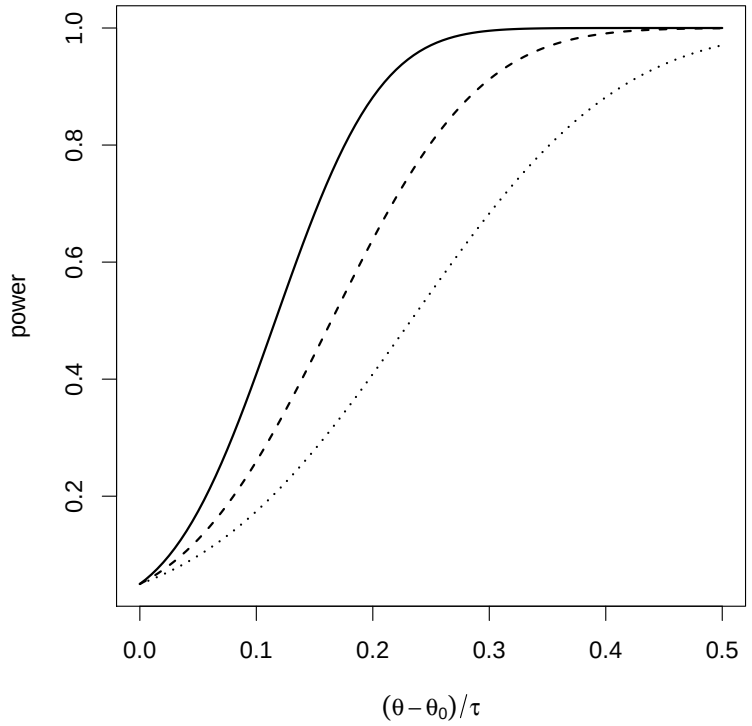
Power (cont.)

Recall that the calculation on slides 182–184 was for an upper-tailed test. Thus the alternatives of interest are for $\theta > \theta_0$. Hence we rewrite this equation again

$$\Pr_\theta(T \geq z_\alpha) \approx \Pr_{\theta_0}\left(T \geq z_\alpha - \frac{\theta - \theta_0}{\tau/\sqrt{n}}\right)$$

replacing t by z_α and $\theta_0 - \theta$ by $-(\theta - \theta_0)$. This, considered as a function of θ is the *power function* of the upper-tailed test. It depends on the hypothetical value of the nuisance parameter τ and the sample size. Alternatively, we could consider it a function of the standardized treatment effect $(\theta - \theta_0)/\tau$ and the sample size.

Power (cont.)



Power curves for upper-tailed z test and $\alpha = 0.05$: solid line is $n = 200$, dashed line is $n = 100$, dotted line is $n = 50$.

Power (cont.)

- Power increases from α to 1 as standardized effect $(\theta - \theta_0)/\tau$ increases from zero to infinity.

- Power increases from α to 1 as sample size increases from zero to infinity.

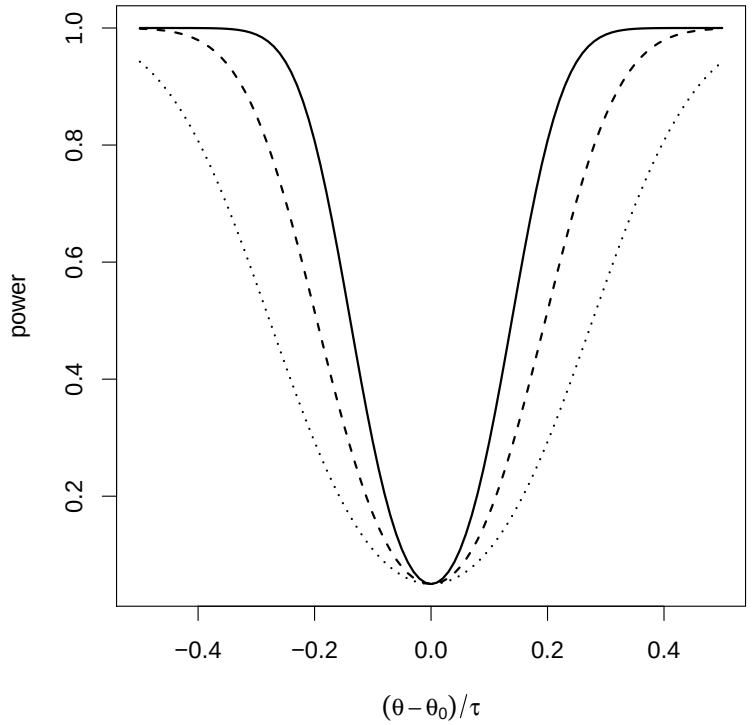
The first is not under control of the experimenters. The effect size is what it is, and although hypothetical in the power calculation, should be realistic. The second is under control of the experimenters. The sample size should be chosen so that the power will be reasonably large.

Power (cont.)

Two-tailed tests are similar

$$\begin{aligned}\Pr_\theta(|T| \geq z_{\alpha/2}) \\ &\approx \Pr_{\theta_0}\left(T \geq z_{\alpha/2} - \frac{\theta - \theta_0}{\tau/\sqrt{n}}\right) + \Pr_{\theta_0}\left(T \leq -z_{\alpha/2} - \frac{\theta - \theta_0}{\tau/\sqrt{n}}\right)\end{aligned}$$

Power (cont.)



Power curves for two-tailed z test and $\alpha = 0.05$: solid line is n = 200, dashed line is $n = 100$, dotted line is $n = 50$.

Power (cont.)

Power calculations are more complicated when the reference distribution is not normal. When the reference distribution is t, F, or chi-square, then the distribution under the alternative hypothesis is so-called noncentral t, F, or chi-square, respectively. And these are new distributions, not on the brand name distributions handout.

R can calculate for these distributions. The R functions `pt`, `pf`, and `pchisq` have a noncentrality parameter argument `ncp` that when supplied calculates using the noncentral distribution.

We will only look at noncentral t here.

Power (cont.)

If Z is standard normal and Y is chi$^2(\nu)$ and Z and Y are independent, then

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

has the $t(\nu)$ distribution. Now we define

$$T = \frac{Z + \delta}{\sqrt{Y/\nu}}$$

to have the noncentral t distribution with degrees of freedom ν and noncentrality parameter δ, denoted $t(\nu, \delta)$.

Power (cont.)

Now we repeat our derivation of the power curve, this time for t tests. The asymptotically pivotal quantity is

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

This is of the form $Z/\sqrt{Y/\nu}$ where $\nu = n - 1$ and

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

$$Y = \frac{(n-1)S_n^2}{\sigma^2}$$

Power (cont.)

The test statistic is

$$\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}$$

When the true unknown parameter value is μ, the numerator of the T

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

is normal but not standard normal, since

$$E(Z) = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

$$\text{var}(Z) = 1$$

Power (cont.)

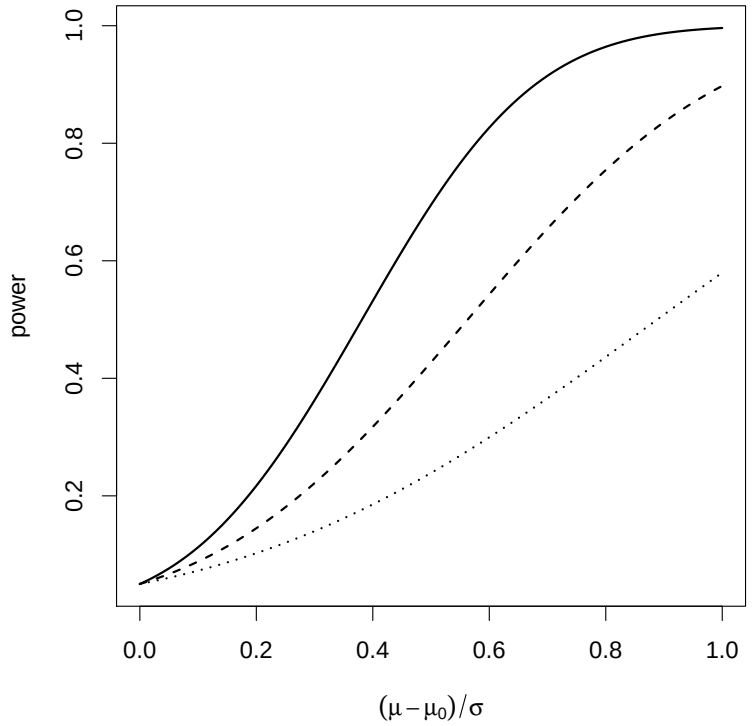
Thus

$$Z = Z^* + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

where Z^* is standard normal and T has the $t(n-1, \delta)$ distribution with

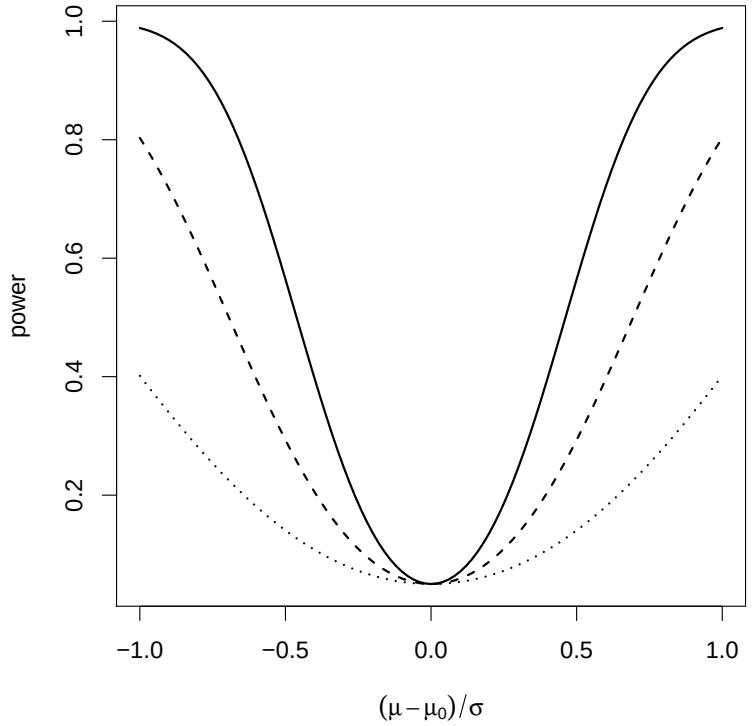
$$\delta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

Power (cont.)



Power curves for upper-tailed t test and $\alpha = 0.05$: solid line is $n = 20$, dashed line is $n = 10$, dotted line is $n = 5$.

Power (cont.)



Power curves for two-tailed t test and $\alpha = 0.05$: solid line is $n = 20$, dashed line is $n = 10$, dotted line is $n = 5$.

Power (cont.)

Qualitatively, we have the same behavior; the power increases from α to 1 as either the standardized effect size increases or the sample size increases. The power curves look much the same whether the normal distribution or the noncentral t distribution is used.

The details of the calculations differ. See computer examples web page.

Correction for Multiple Testing

One way to correct for multiple testing is to consider the multiple tests one combined test and control the level of the combined test, which is called the *familywise error rate* (FWER). Again the decision-theoretic viewpoint makes the theory simpler.

Let us say the combined test rejects H_0 if any one of the separate tests rejects its own particular H_0. Then

$$\begin{aligned}\text{Pr}(\text{combined test rejects } H_0) &= \text{Pr}\left(\bigcup_{j=1}^{k} \text{test } j \text{ rejects its } H_0\right) \\ &\leq \sum_{j=1}^{k} \text{Pr}(\text{test } j \text{ rejects its } H_0)\end{aligned}$$

by subadditivity of probability (5101 deck 2, slide 138).

Correction for Multiple Testing (cont.)

Hence if we make

$$\text{Pr}(\text{test } j \text{ rejects its } H_0) = \frac{\alpha}{k}, \qquad \text{for all } j$$

then the combined test will have significance level less than or equal to α.

The P-value of the combined test is then the smallest α for which the combined test rejects H_0, which is k times the smallest α for which one of the multiple tests rejects. Hence the P-value for the combined test formed from k multiple tests is just k times the smallest P-value for any of the multiple tests.

This is known as *Bonferroni correction*.

Correction for Multiple Testing (cont.)

A P-value $P = 0.01$ looks highly statistically significant before we find out that $k = 6$ tests were done, and the Bonferroni corrected P-value is $P = 0.06$.

Many scientists do not like Bonferroni correction, because it makes $P < 0.05$ much harder to obtain. Also they complain that Bonferroni is too conservative. It just provides a bound, not an exact correction.

Thus multiple testing without correction is often done. But it should not be.

Duality of Tests and Confidence Intervals

If we take the decision-theoretic view of hypothesis tests, then there is a duality between exact $100(1-\alpha)\%$ two-sided confidence intervals for a parameter θ and tests of the hypotheses

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

having significance level α.

The test accepts H_0 if and only if the confidence interval contains θ_0.

The confidence interval consists of all real numbers θ_0 for which the test accepts H_0.

Correction for Multiple Testing (cont.)

One application of the duality of tests and confidence intervals is that the dogma clearly applies to confidence intervals too.

Do only one confidence interval. Otherwise you must correct to obtain simultaneous coverage.

The Bonferroni correction for confidence intervals uses confidence level $100(1 - \alpha/k)\%$ for each of k intervals.

If one does not do such correction, then it must be clearly stated that the stated confidence level is not for simultaneous coverage.

Sign Test and Related Procedures

So far we have the asymptotic normal distribution of the sample median

$$\widetilde{X}_n \approx \mathcal{N}\left(\mu, \frac{1}{4nf(\mu)^2}\right)$$

where μ is the population median and f is the PDF of the true unknown distribution of the data.

If we assume an IID normal sample, then the asymptotic variance is $\pi\sigma^2/2n$ so a $100(1-\alpha)\%$ confidence interval for the population median using the plug-in principle would be

$$\widetilde{X}_n \pm z_{\alpha/2} S_n \sqrt{\frac{\pi}{2n}}$$

But now we seek a nonparametric procedure that does not assume any particular parametric model for the data.

Sign Test and Related Procedures (cont.)

We start with a hypothesis test. Suppose X_1, X_2, ..., X_n are IID from a continuous distribution having population median θ. We consider one-tailed and two-tailed tests with null hypothesis

$$H_0 : \theta = \theta_0$$

and test statistic

$$T = \sum_{i=1}^{n} I_{(\theta_0, \infty)}(X_i)$$

(the number of X_i greater than θ_0). Under H_0, the distribution of T is Bin$(n, 1/2)$.

Sign Test and Related Procedures (cont.)

Knowing the distribution of the test statistic under H_0, we make P-values in the usual way.

$\Pr(T \geq t)$ is the P-value of the upper-tailed test.

$\Pr(T \leq t)$ is the P-value of the lower-tailed test.

If $t \geq n/2$, then $\Pr(T \leq n - t \text{ or } T \geq t)$ is the P-value of the two-tailed test.

If $t \leq n/2$, then $\Pr(T \leq t \text{ or } T \geq n - t)$ is the P-value of the two-tailed test.

Sign Test and Related Procedures (cont.)

The only thing that is a bit tricky is the distribution of the test statistic is symmetric but the center of symmetry is $n/2$ rather than zero.

Hence $|T|$ is not a sensible test statistic for the two-tailed test. Rather, we reject H_0 when $|T - n/2|$ is large.

Sign Test and Related Procedures (cont.)

Another thing that is different from procedures with a continuous test statistic is that, in the decision-theoretic view, only a few significance levels are exactly achievable. If $n = 10$, then

```
Rweb> round(pbinom(0:4, 10, 1 / 2), 5)
[1] 0.00098 0.01074 0.05469 0.17188 0.37695
```

are the only numbers below 1/2 that can be either P-values or significance levels of one-tailed tests, and

```
Rweb> round(2 * pbinom(0:3, 10, 1 / 2), 5)
[1] 0.00195 0.02148 0.10937 0.34375
```

are the only numbers below 1/2 that can be either P-values or significance levels of two-tailed tests.

Sign Test and Related Procedures (cont.)

As usual let $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$ denote the order statistics for the sample.

Select $\alpha < 1$ and $k \geq 0$ such that $\Pr(T \leq k) = \alpha/2$. Then $X_{(k+1)} < \theta < X_{(n-k)}$ is an exact $100(1-\alpha)\%$ confidence interval for the population median.

Proof on following slide.

Sign Test and Related Procedures (cont.)

The sign test accepts $H_0 : \theta = \theta_0$ at level α when $k < T < n - k$. This is the same as saying at least $k + 1$ of the data points are below θ_0 and at least $k + 1$ of the data points are above θ_0. And this is the same as saying $X_{(k+1)} < \theta_0 < X_{(n-k)}$.

Only the last bit is tricky. At least $k + 1$ of the data points are above θ_0 if and only if the $k + 1$ largest of the order statistics are above θ_0, and these are

$$X_{(n-i)}, \qquad i = 0, \ldots, k$$

Sign Test and Related Procedures (cont.)

Given a nested family of confidence intervals, one for each α, like t confidence intervals or these intervals associated with the sign test, the point estimate arrived at by making α nearly one-half so the interval shrinks to nearly zero length is called the *Hodges-Lehmann estimator* associated with the procedure.

For the t confidence intervals, the Hodges-Lehmann estimator is the sample mean.

For the confidence intervals associated with the sign test, the Hodges-Lehmann estimator is the sample median.

Sign Test and Related Procedures (cont.)

Thus procedures come in trios: hypothesis test, confidence interval, and point estimate. If we have a test about a location parameter, then we also have the confidence interval dual to the test, and we also have the Hodges-Lehmann estimator obtained from the confidence interval.

Fuzzy Tests and Confidence Intervals

It is both a theoretical and a practical nuisance that the discreteness of the distribution of the test statistic under H_0 means only a few significance levels or confidence levels are possible. And the conventional 0.05 significance level or 95% confidence level cannot be chosen.

This means procedures based on a discrete test statistic are not comparable to those with a continuous test statistic.

To make them comparable, the notion of randomized tests was invented.

Fuzzy Tests and Confidence Intervals (cont.)

Suppose first we are doing a lower-tailed sign test. The conventional P-value is $\Pr(T \le t)$. The randomized test has P-value that is uniformly distributed on the interval

$$\Big(\Pr(T < t), \Pr(T \le t)\Big)$$

Note that the randomness in the P-value is artificial. It is not related to the randomness in the random sample. Rather it is introduced by the statistician as a mathematical trick.

All probability calculations involve both kinds of randomness: artificial and from random sampling.

Fuzzy Tests and Confidence Intervals (cont.)

For any α between $\Pr(T < k)$ and $\Pr(T \leq k)$, the probability that $P < \alpha$ so the test rejects H_0 is

$$\Pr(T < k) + \Pr(T = k) \cdot \frac{\alpha - \Pr(T < k)}{\Pr(T \leq k) - \Pr(T < k)} = \alpha$$

so the significance level of the test is α.

Fuzzy Tests and Confidence Intervals (cont.)

Two statisticians can analyze the same data using the same randomized test procedure and get different results due to the artificial randomization.

This property is so absurd that randomized procedures are never used in actual applications. Fuzzy P-values were introduced to fix this problem.

Fuzzy procedures are the same as randomized procedures except that the randomization is not done only described. The P-value is reported to be a random variable uniformly distributed on the interval

$$\Big(\mathsf{Pr}(T < t), \mathsf{Pr}(T \leq t)\Big)$$

that is, rather than a single number, the interval is reported.

Fuzzy Tests and Confidence Intervals (cont.)

Fuzzy P-values are comparable to ordinary P-values for procedures having a continuous distribution of the test statistic under H_0.

Consider a sign test with sample size $n = 10$. We saw that the possible P-values for the nonrandomized test were

```
Rweb> round(pbinom(0:4, 10, 1 / 2), 5)
[1] 0.00098 0.01074 0.05469 0.17188 0.37695
```

Fuzzy Tests and Confidence Intervals (cont.)

If one observes $t = 2$, then one reports $P = 0.055$. Not statistically significant according to worshipers of the number 0.05.

But the next smallest possible P-value is $P = 0.011$ corresponding to $t = 1$. So this is not analogous to a t-test. Intuitions that that come from experience with t-tests are not transferable.

For the fuzzy test, if one observes $t = 2$, then one reports P is uniformly distributed on the interval $(0.011, 0.055)$. It is mostly below 0.05, hence this result is more analogous to a t-test with $P < 0.05$ than one with $P > 0.05$.

Fuzzy Tests and Confidence Intervals (cont.)

Another way to interpret the fuzzy P-value is to consider what the randomized test would do if it were done. If the fuzzy P-value is uniformly distributed on the interval $(0.011, 0.055)$, then a randomized test would reject H_0 at level $\alpha = 0.05$ if $P < \alpha$, which happens with probability

$$\frac{0.05 - 0.011}{0.055 - 0.011} = 0.886$$

Thus if many statisticians did the randomized test for this same data, not all would reject H_0 but 88.6% of them would.

These data are more statistically significant than the conventional $P = 0.055$ suggests.

Fuzzy Tests and Confidence Intervals (cont.)

For a two-tailed sign test, the fuzzy P-value is uniform on the interval

$$\Big(2\,\mathsf{Pr}(T < t), 2\,\mathsf{Pr}(T \leq t)\Big)$$

in case $t < n/2$ and uniform on the interval

$$\Big(2\,\mathsf{Pr}(T \geq t), 2\,\mathsf{Pr}(T > t)\Big)$$

in case $t > n/2$ and uniform on the interval

$$\Big(\mathsf{Pr}(T \neq n/2), 1\Big)$$

in case $t = n/2$.

Fuzzy Tests and Confidence Intervals (cont.)

The corresponding fuzzy or randomized confidence intervals find k such that

$$\Pr(T < k) \leq \frac{\alpha}{2} \leq \Pr(T \leq k)$$

the randomized confidence interval is $X_{(k+1)} < \theta < X_{(n-k)}$ with probability

$$p = \frac{\alpha - \Pr(T < k)}{\Pr(T \leq k) - \Pr(T < k)}$$

and is $X_{(k)} < \theta < X_{(n-k+1)}$ with probability $1 - p$, where by convention $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$.

The fuzzy confidence interval reports the two intervals and their probabilities p and $1 - p$.

Signed Rank Test and Related Procedures

Suppose X_1, X_2, ..., X_n are IID from a continuous symmetric distribution having population center of symmetry θ. We consider one-tailed and two-tailed tests with null hypothesis

$$H_0 : \theta = \theta_0$$

Let

$$Y_i = |X_i - \theta_0|$$

and let R_i be the rank of Y_i in the sorted order, that is, $Y_i = Y_{(R_i)}$. Let Z_i be R_i times the sign of $X_i - \theta_0$. The Z_i are called the *signed ranks*.

Because of the assumption of continuity, no ties are possible among the X_i, either with each other or with θ_0. The ranks and signs are unambiguously determined with probability one.

Signed Rank Test and Related Procedures (cont.)

Let

$$T = \sum_{i=1}^{n} Z_i I_{(0,\infty)}(Z_i)$$

(the sum of the positive signed ranks). Under H_0, the distribution of T is called the distribution of the signed rank statistic, calculated by the R functions `psignrank`, `qsignrank`, etc.

When H_0 is true, conditional on the values of the Y_i, the signs of the Z_i are IID and $+$ and $-$ are equally probable. This makes the distribution of T symmetric about the midpoint of its range. The smallest possible value is zero, the largest possible value is $n(n+1)/2$, and the midpoint is $n(n+1)/4$.

Signed Rank Test and Related Procedures (cont.)

Knowing the distribution of the test statistic under H_0, we make P-values in the usual way. This slide is an exact copy of slide 210 except that $N = n(n+1)/2$ replaces n.

$\Pr(T \geq t)$ is the P-value of the upper-tailed test.

$\Pr(T \leq t)$ is the P-value of the lower-tailed test.

If $t \geq N/2$, then $\Pr(T \leq N - t \text{ or } T \geq t)$ is the P-value of the two-tailed test.

If $t \leq N/2$, then $\Pr(T \leq t \text{ or } T \geq N - t)$ is the P-value of the two-tailed test.

Signed Rank Test and Related Procedures (cont.)

The only thing that is a bit tricky is the distribution of the test statistic is symmetric but the center of symmetry is $n(n+1)/4$ rather than zero.

Hence $|T|$ is not a sensible test statistic for the two-tailed test. Rather, we reject H_0 when $|T - n(n+1)/4|$ is large.

Signed Rank Test and Related Procedures (cont.)

As with the sign test, in the decision-theoretic view, only a few significance levels are exactly achievable. However, more are achievable than for a sign test with the same sample size. If $n = 10$, then

```
Rweb> round(psignrank(0:15, 10), 5)
 [1] 0.00098 0.00195 0.00293 0.00488 0.00684 0.00977 0.01367
 [8] 0.01855 0.02441 0.03223 0.04199 0.05273 0.06543 0.08008
[15] 0.09668 0.11621
```

are the only numbers below 0.12 that can be either P-values or significance levels of one-tailed tests.

Signed Rank Test and Related Procedures (cont.)

Similarly,

```
Rweb> round(2 * psignrank(0:11, 10), 5)
 [1] 0.00195 0.00391 0.00586 0.00977 0.01367 0.01953 0.02734
 [8] 0.03711 0.04883 0.06445 0.08398 0.10547
```

are the only numbers below 0.12 that can be either P-values or significance levels of two-tailed tests.

Signed Rank Test and Related Procedures (cont.)

To find the confidence interval dual to the signed rank test, we need to define the Walsh averages, which are the $n(n+1)/2$ numbers

$$\frac{X_i + X_j}{2}, \qquad i \leq j$$

In case $i = j$, the Walsh average is just X_i. Otherwise, it is the number halfway between X_i and X_j. There are n Walsh averages of the first kind and $n(n-1)/2$ of the second kind.

Claim: the test statistic T is equal to the number of Walsh averages greater than θ_0, call that T^*.

Signed Rank Test and Related Procedures (cont.)

The proof uses mathematical induction. Consider the data fixed, just a set of numbers. Because of the continuity assumption all of the Walsh averages are different with probability one.

If θ_0 is greater than all of the data points, then all of the signed ranks are negative and all of the Walsh averages are less than θ_0. Hence $T = T^* = 0$. That is the base of the induction.

As θ_0 moves from above all Walsh averages to below them, neither T nor T^* changes except when θ_0 passes a Walsh average, in which case both increase by one. That is the induction step.

When we have proved the induction step, that proves $T = T^*$ regardless of the value of θ_0.

Signed Rank Test and Related Procedures (cont.)

Clearly T^* increases by one each time θ_0 passes a Walsh average going from above to below. We only need verify that the same goes for T.

Induction step at Walsh average $W = X_i$. For θ_0 near X_i we have $R_i = 1$. As θ_0 moves from above X_i to below it, Z_i changes from $-$ to $+$. None of the other Z_j change. Hence T increases by one.

Induction step at Walsh average $W = (X_i + X_j)/2$. Say $X_i < X_j$. For θ_0 near W we have R_i and R_j with consecutive ranks because $R_j - W \approx W - R_i$. As θ_0 moves from above W to below it, R_j and R_i swap values and R_j increases by one. Since $Z_j > 0$ and $Z_i < 0$, this increases T^* by one.

Signed Rank Test and Related Procedures (cont.)

Let $W_{(1)}, W_{(2)}, \ldots, W_{(N)}$ denote the order statistics of the Walsh averages, where $N = n(n+1)/2$.

Select $\alpha < 1$ and $k \geq 0$ such that $\Pr(T \leq k) = \alpha/2$. Then $W_{(k+1)} < \theta < W_{(N-k)}$ is an exact $100(1-\alpha)\%$ confidence interval for the population center of symmetry.

The proof is exactly like the proof for the confidence interval associated with the sign test: Walsh averages replace the data, `psignrank` replaces `pbinom`, and $N = n(n+1)/2$ replaces n, but everything else remains the same.

Signed Rank Test and Related Procedures (cont.)

The Hodges-Lehmann estimator associated with the signed rank test is the median of the Walsh averages.

There are fuzzy hypothesis tests and confidence intervals for the signed rank test done by the `fuzzyRankTests` package. We won't spend time on them because they are very similar to the ones for the sign test.

Asymptotic Relative Efficiency Revisited

Asymptotic relative efficiency (ARE) for hypothesis tests and confidence intervals is defined to be the same as the ARE for the corresponding point estimators.

We know that the ARE of the sample median versus the sample mean as point estimators (and consequently the ARE of the sign test versus the t test) can vary from zero to infinity. For distributions not having second moments the sample median and the sign test are infinitely better than the sample mean and the t test. For continuous distributions having PDF equal to zero at the median, the reverse holds: the sample median and the sign test are infinitely worse than the sample mean and the t test.

Asymptotic Relative Efficiency Revisited (cont.)

We shall see (deck 3) that when the population distribution is exactly normal, the sample mean and the t test are the best of all possible procedures.

How much do we lose in this case if we use the sign test or the signed rank test and their related confidence intervals and point estimators?

Asymptotic Relative Efficiency Revisited (cont.)

We calculated (slide 53) that, if the population distribution is normal, then the ARE of the sample median relative to the sample mean is $2/\pi \approx 0.6366$. So we lose a lot.

It can be shown (but we won't) that, if the population is normal, then the ARE of the median of the sample Walsh averages relative to the sample mean is

$$12 \left(\int_{-\infty}^{\infty} f(x)^2 \, dx \right)^2 = \frac{3}{\pi} \approx 0.955$$

where f is the PDF of the standard normal distribution. So we lose very little.