SML 201 – Week 7

John D. Storey

Spring 2016

Contents

Central Limit Theorem

Linear Transformation of a RV

Suppose that X is a random variable and that a and b are constants. Then:

$$\mathrm{E}\left[a + bX\right] = a + b\mathrm{E}[X]$$

$$\mathrm{Var}\left(a + bX\right) = b^2\mathrm{Var}(X)$$

Sums of Random Variables

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) random variables, then:

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}[X_i]$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Means of Random Variables

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) random variables. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be their sample mean. Then:

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}[X_i]$$

$$\mathrm{Var}\left(\overline{X}\right) = \frac{1}{n}\mathrm{Var}(X_i)$$

Statement of the CLT

Suppose $X_1, X_2, \ldots, X_n$ are iid rv's with population mean $\mathrm{E}[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$.

Then for "large n", $\sqrt{n}(\overline{X} - \mu)$ approximately follows the Normal(0, σ^2) distribution.

As $n \to \infty$, this approximation becomes exact.

Example: Calculations

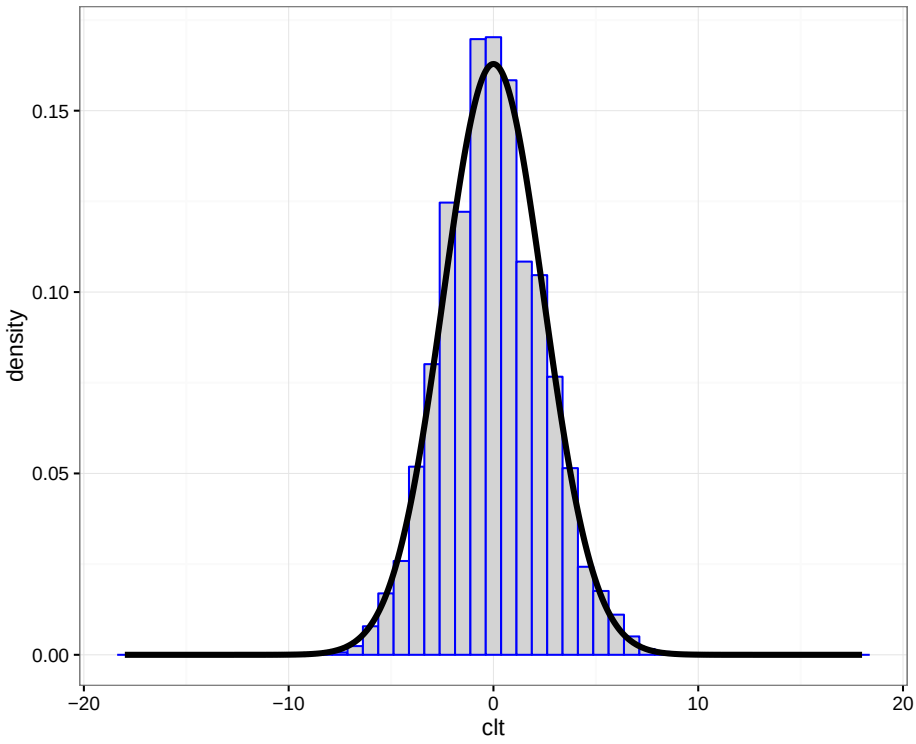
Let $X_1, X_2, \ldots, X_{40}$ be iid Poisson(λ) with $\lambda = 6$.

We will form $\sqrt{40}(\overline{X} - 6)$ over 10,000 realizations and compare their distribution to a Normal(0, 6) distribution.

```
> x <- replicate(n=1e4, expr=rpois(n=40, lambda=6),
+                simplify="matrix")
> x_bar <- apply(x, 2, mean)
> clt <- sqrt(40)*(x_bar - 6)
>
> df <- data.frame(clt=clt, x = seq(-18,18,length.out=1e4),
+                  y = dnorm(seq(-18,18,length.out=1e4),
+                            sd=sqrt(6)))
```

Example: Plot

```
> ggplot(data=df) +
+   geom_histogram(aes(x=clt, y=..density..), color="blue",
+                  fill="lightgray", binwidth=0.75) +
+   geom_line(aes(x=x, y=y), size=1.5)
```



Statistical Inference

Data Collection as a Probability

- Suppose data are collected in such a way that it is randomly observed according to a probability distribution
- If that probability distribution can be parameterized, then it is possible that the parameters describe key characteristics of the population of interest
- **Statistical inference** reverse engineers this process to estimate the unknown values of the parameters and express a measure of uncertainty about these estimates

Example: Simple Random Sample

Individuals are uniformly and independently randomly sampled from a population.

The measurements taken on these individuals are then modeled as random variables, specifically random realizations from the complete population of

values.

Simple random samples form the basis of modern surveys.

Example: Randomized Controlled Trial

Individuals under study are randomly assigned to one of two or more available treatments.

This induces randomization directly into the study and breaks the relationship between the treatments and other variables that may be influencing the response of interest.

This is the gold standard study design in clinical trials to assess the evidence that a new drug works on a given disease.

Parameters and Statistics

- A **parameter** is a number that describes a population
  - A parameter is often a fixed number
  - We usually do not know its value

- A **statistic** is a number calculated from a sample of data
- A statistic is used to estimate a parameter

Sampling Distribution

The **sampling distribution** of a statistic is the probability disribution of the statistic under repeated realizations of the data from the assumed data generating probability distribution.

The sampling distribution is how we connect an observed statistic to the population.

Example: Fair Coin?

Suppose I claim that a specific coin is fair, i.e., that it lands on heads or tails with equal probability.

I flip it 20 times and it lands on heads 16 times.

1. My data is $x = 16$ heads out of $n = 20$ flips.
2. My data generation model is $X \sim \text{Binomial}(20, p)$.
3. I form the statistic $\hat{p} = 16/20$ as an estimate of p.

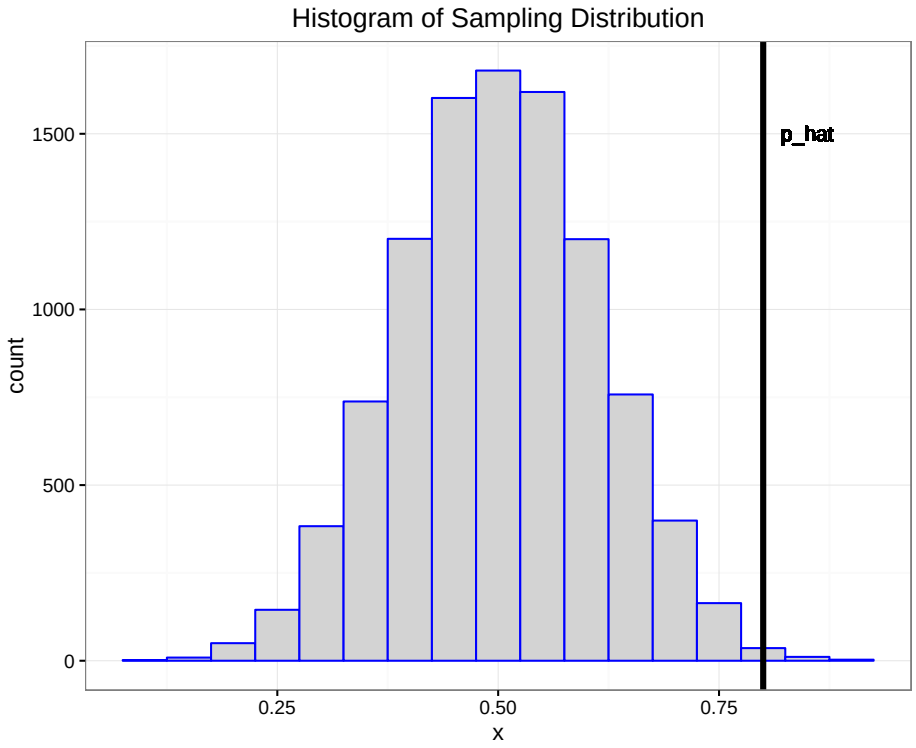
Example (cont'd)

Let's simulate 10,000 times what my estimate would look like if $p = 0.5$ and I repeated the 20 coin flips over and over.

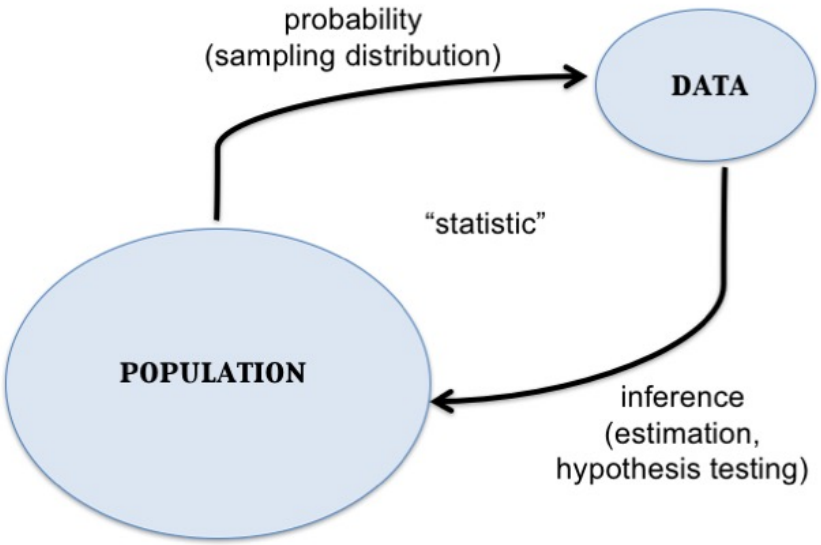
```
> x <- replicate(n=1e4, expr=rbinom(1, size=20, prob=0.5))
> sim_p_hat <- x/20
> my_p_hat <- 16/20
```

What can I do with this information?

Example (cont'd)



Central Dogma of Inference



Inference Goals and Strategies

Basic Idea

Data are collected in such a way that there exists a reasonable probability model for this process that involves parameters informative about the population.

Common Goals:

1. Form point estimates the parameters
2. Quantify uncertainty on the estimates
3. Test hypotheses on the parameters

Normal Example

Suppose a simple random sample of n data points is collected so that the following model of the data is reasonable: $X_1, X_2, \ldots, X_n$ are iid Normal(μ, σ^2).

The goal is to do inference on μ, the population mean.

For simplicity, assume that σ^2 is known (e.g., $\sigma^2 = 1$).

Point Estimate of μ

There are a number of ways to form an estimate of μ, but one that has several justifications is the sample mean:

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where $x_1, x_2, \ldots, x_n$ are the observed data points.

Sampling Distribution of $\hat{\mu}$

If we were to repeat this study over and over, how would $\hat{\mu}$ behave?

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\overline{X} \sim \text{Normal}(\mu, \sigma^2/n)$$

How do we use this to quantify uncertainty and test hypotheses?

Pivotal Statistic

One *very useful* strategy is to work backwards from a pivotal statistic, which is a statistic that does not depend on any unknown paramaters.

Example:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

Note that in general for a rv Y it is the case that $(Y - \text{E}[Y])/\sqrt{\text{Var}(Y)}$ has population mean 0 and variance 1.

Confidence Intervals

Goal

Once we have a point estimate of a parameter, we would like a measure of its uncertainty.

Given that we are working within a probabilistic framework, the natural language of uncertainty is through probability statements.

We interpret this measure of uncertainty in terms of hypothetical repetitions of the sampling scheme we used to collect the original data set.

Formulation

Confidence intervals take the form

$$(\hat{\mu} - C_\ell, \hat{\mu} + C_u)$$

where

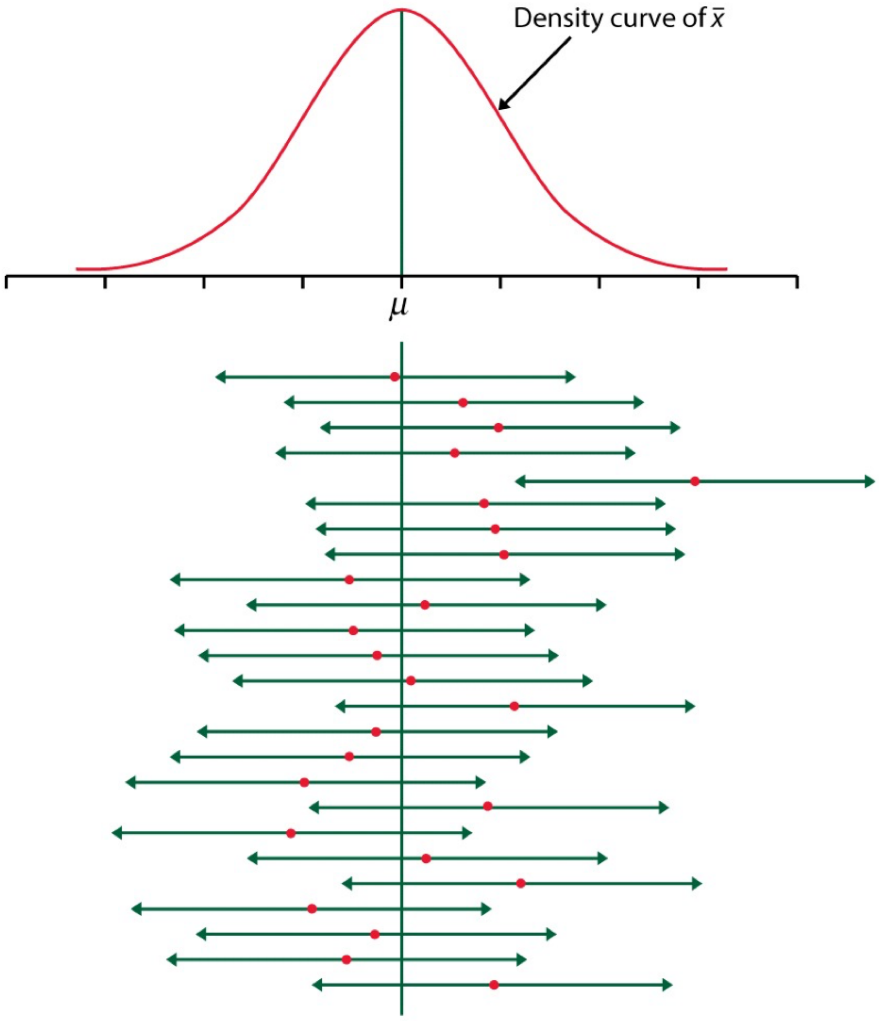
$$\Pr(\mu - C_\ell \leq \hat{\mu} \leq \mu + C_u)$$

forms the "level" or coverage probability of the interval.

Interpretation

If we repeat the study many times, then the CI $(\hat{\mu} - C_\ell, \hat{\mu} + C_u)$ will contain the true value μ with a long run frequency equal to $\Pr(\mu - C_\ell \leq \hat{\mu} \leq \mu + C_u)$.

A CI calculated on an observed data set is *not* intepreted as: "There is probability $\Pr(\mu - C_\ell \leq \hat{\mu} \leq \mu + C_u)$ that μ is in our calculated $(\hat{\mu} - C_\ell, \hat{\mu} + C_u)$." Why not?



A Normal CI

If $Z \sim$ Normal(0,1), then $\Pr(-1.96 \leq Z \leq 1.96) = 0.95$.

$$
\begin{aligned}
0.95 &= \Pr\left(-1.96 \leq \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) && (1)\\
&= \Pr\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \hat{\mu} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) && (2)\\
&= \Pr\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \hat{\mu} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) && (3)
\end{aligned}
$$

Therefore, $\left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ forms a 95% confidence interval of μ.

A Simulation

```
> mu <- 5
> n <- 20
> x <- replicate(10000, rnorm(n=n, mean=mu)) # 10000 studies
> m <- apply(x, 2, mean) # the estimate for each study
> ci <- cbind(m - 1.96/sqrt(n), m + 1.96/sqrt(n))
> head(ci)
         [,1]     [,2]
[1,] 4.613983 5.490522
[2,] 4.718898 5.595437
[3,] 4.857944 5.734483
[4,] 4.697341 5.573880
[5,] 4.621864 5.498403
[6,] 4.494349 5.370888
```

```
> cover <- (mu > ci[,1]) & (mu < ci[,2])
> mean(cover)
[1] 0.9487
```

Normal$(0, 1)$ Percentiles

Above we constructed a 95% CI. How do we construct (1-α)-level CIs?

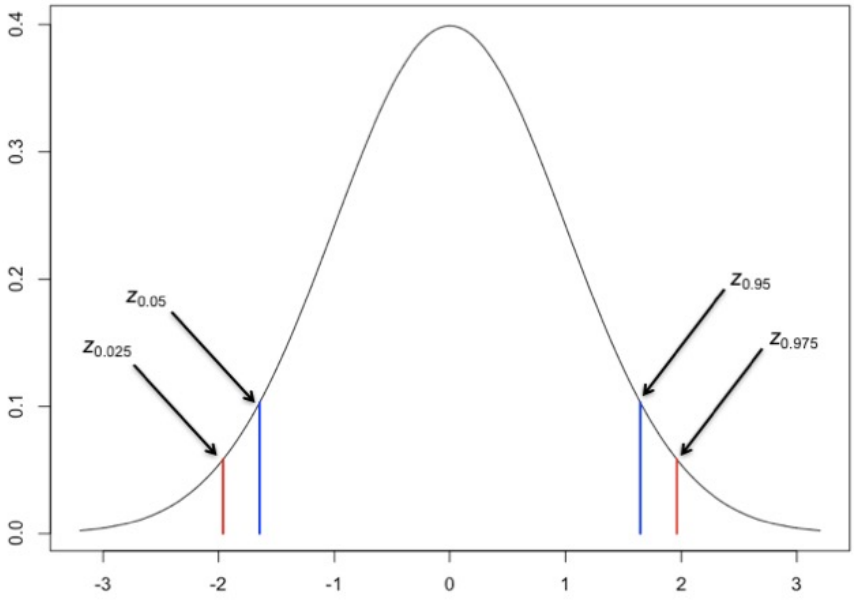
Let z_α be the α percentile of the Normal(0,1) distribution.

If $Z \sim$ Normal(0,1), then

$$\begin{aligned} 1 - \alpha &= \Pr(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) \\ &= \Pr(-|z_{\alpha/2}| \leq Z \leq |z_{\alpha/2}|) \end{aligned}$$

```
> qnorm(0.025)
[1] -1.959964
> qnorm(0.975)
[1] 1.959964
```

Commonly Used Percentiles



$(1-\alpha)$-Level CIs

If $Z \sim$ Normal(0,1), then $\Pr(-|z_{\alpha/2}| \leq Z \leq |z_{\alpha/2}|) = 1-\alpha$.

Repeating the steps from the 95% CI case, we get the following is a $(1-\alpha)$-Level CI for μ:

$$\left(\hat{\mu} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}, \hat{\mu} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}\right)$$

One-Sided CIs

The CIs we have considered so far are "two-sided". Sometimes we are also interested in "one-sided" CIs.

If $Z \sim$ Normal(0,1), then $1-\alpha = \Pr(Z \geq -|z_\alpha|)$ and $1-\alpha = \Pr(Z \leq |z_\alpha|)$. We can use this fact along with the earlier derivations to show that the following are valid CIs:

$$(1-\alpha)\text{-level upper: } \left(-\infty, \hat{\mu} + |z_\alpha|\frac{\sigma}{\sqrt{n}}\right)$$

$(1-\alpha)$-level lower: $\left(\hat{\mu} - |z_\alpha|\frac{\sigma}{\sqrt{n}}, \infty\right)$

Hypothesis Tests

Example: HT on Fairness of a Coin

Suppose I claim that a specific coin is fair, i.e., that it lands on heads or tails with equal probability.

I flip it 20 times and it lands on heads 16 times.

1. My data is $x = 16$ heads out of $n = 20$ flips.
2. My data generation model is $X \sim \text{Binomial}(20, p)$.
3. I form the statistic $\hat{p} = 16/20$ as an estimate of p.

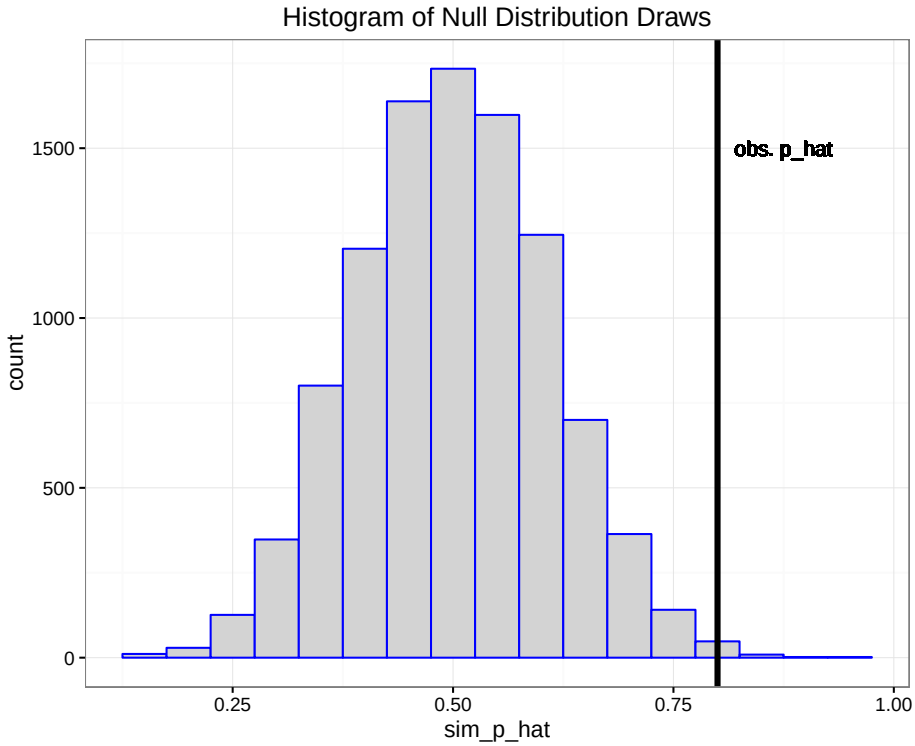
More formally, I want to **test the hypothesis**: $H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$ under the model $X \sim \text{Binomial}(20, p)$ based on the **test statistic** $\hat{p} = X/n$.

Example (cont'd): Null Distribution

Let's simulate 10,000 times what my estimate would look like if $p = 0.5$ and I repeated the 20 coin flips over and over.

```
> x <- replicate(n=1e4, expr=rbinom(1, size=20, prob=0.5))
> sim_p_hat <- x/20
> my_p_hat <- 16/20
```

The vector `sim_p_hat` contains 10,000 draws from the **null distribution**, i.e., the distribution of my test statstic $\hat{p} = X/n$ when $H_0 : p = 0.5$ is true.



Example (cont'd): P-value

The deviation of the test statistic from the null hypothesis can be measured by $|\hat{p} - 0.5|$.

Let's compare our observed deviation $|16/20 - 0.5|$ to the 10,000 simulated null data sets. Specifically, let's calculate the frequency by which these 10,000 cases are **as or more extreme** than the observed test statistic.

```
> sum(abs(sim_p_hat-0.5) >= abs(my_p_hat-0.5))/1e4
[1] 0.0072
```

This quantity is called the **p-value** of the hypothesis test.

A Caveat

This example is a simplification of a more general framework for testing statistical hypotheses.

Given the intuition provided by the example, let's now formalize these ideas.

Definition

- A **hypothesis test** or **significance test** is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess
- The results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree
- The **null hypothesis** (H_0) is the statement being tested, typically the status quo
- The **alternative hypothesis** (H_1) is the complement of the null, and it is often the "interesting" state

Return to Normal Example

Let's return to our Normal example in order to demonstrate the framework.

Suppose a simple random sample of n data points is collected so that the following model of the data is reasonable: $X_1, X_2, \ldots, X_n$ are iid Normal(μ, σ^2).

The goal is to do test a hypothesis on μ, the population mean.

For simplicity, assume that σ^2 is known (e.g., $\sigma^2 = 1$).

HTs on Parameter Values

Hypothesis tests are usually formulated in terms of values of parameters. For example:

$$H_0 : \mu = 5$$

$$H_1 : \mu \neq 5$$

Note that the choice of 5 here is arbitrary, for illustrative purposes only. In a typical real world problem, the values that define the hypotheses will be clear from the context.

Two-Sided vs. One-Sided HT

Hypothesis tests can be two-sided or one-sided:

$$H_0 : \mu = 5 \text{ vs. } H_1 : \mu \neq 5 \text{ (two-sided)}$$

$$H_0 : \mu \leq 5 \text{ vs. } H_1 : \mu > 5 \text{ (one-sided)}$$

$$H_0 : \mu \geq 5 \text{ vs. } H_1 : \mu < 5 \text{ (one-sided)}$$

Test Statistic

A **test statistic** is designed to *quantify the evidence against the null hypothesis in favor of the alternative.* They are usually defined (and justified using math theory) so that the larger the test statistic is, the more evidence there is.

For the Normal example and the two-sided hypothesis ($H_0 : \mu = 5$ vs. $H_1 : \mu \neq 5$), here is our test statistic:

$$|z| = \frac{|\overline{x} - 5|}{\sigma/\sqrt{n}}$$

What would the test statistic be for the one-sided hypothesis tests?

Null Distribution (Two-Sided)

The **null distribution** is the sampling distribution of the test statistic when H_0 is true.

We saw earlier that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$.

When H_0 is true, then $\mu = 5$. So when H_0 is true it follows that

$$Z = \frac{\overline{X} - 5}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

and then probabiliy calculations on $|Z|$ are straightforward. Note that Z is pivotal when H_0 is true!

Null Distribution (One-Sided)

When performing a one-sided hypothesis test, such as $H_0 : \mu \leq 5$ vs. $H_1 : \mu > 5$, the null distribution is typically calculated under the "least favorable" value, which is the boundary value.

In this example it would be $\mu = 5$ and we would again utilize the null distribution

$$Z = \frac{\overline{X} - 5}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

P-values

The **p-value** is defined to be the probability that a test statistic from the null distribution is *as or more extreme than the observed statistic*. In our Normal example on the two-sided hypothesis test, the p-value is

$$\Pr(|Z^*| \geq |z|)$$

where $Z^* \sim \text{Normal}(0, 1)$ and $|z|$ is the value of the test statistic calculated on the data (so it is a fixed number once we observe the data).

Calling a Test "Significant"

A hypothesis test is called **statistically significant** — meaning we reject H_0 in favor of H_1 — if its p-value is sufficiently small.

Commonly used cut-offs are 0.01 or 0.05, although these are not always appropriate.

Applying a specific p-value cut-off to determine significance determines an error rate, which we define next.

Types of Errors

There are two types of errors that can be committed when performing a hypothesis test.

1. A **Type I error** or **false positive** is when a hypothesis test is called signficant and the null hypothesis is actually true.
2. A **Type II error** or **false negative** is when a hypothesis test is not called signficant and the alternative hypothesis is actually true.

Error Rates

- The **Type I error rate** or **false positive rate** is the probability of this type of error given that H_0 is true.
- If a hypothesis test is called significant when p-value $\leq \alpha$ then it has a Type I error rate equal to α.
- The **Type II error rate** or **false negative rate** is the probability of this type of error given that H_1 is true.
- The **power** of a hypothesis test is $1-$ Type II error rate.

Hypothesis tests are usually derived with a goal to control the Type I error rate while maximizing the power.

CLT for Common Estimators

The Normal Example

We formulated both confidence intervals and hypothesis tests under the following "example":

> Suppose a simple random sample of n data points is collected so that the following model of the data is reasonable: $X_1, X_2, \ldots, X_n$ are iid Normal(μ, σ^2). The goal is to do inference on μ, the population mean. For simplicity, assume that σ^2 is known (e.g., $\sigma^2 = 1$).

There is a good reason why we did this.

Normal Pivotal Statistics

The random variable distributions we introduced in Week 6 all have parameter estimators that can be standardized to yield a pivotal statistic with a Normal(0,1) distribution.

For example, if $X \sim \text{Binomial}(n, p)$ and $\hat{p} = X/n$, then for large n it approximately holds that:

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim \text{Normal}(0, 1).$$

We will cover these results next week, which will allow us to directlty leverage the Normal case we worked out this week.

Bayesian Inference

Frequentist Probability

The inference framework we have covered so far uses a **frequentist** intepretation of probability.

We made statements such as, "If we repeat this study over and over, the long run frequency is such that..."

The Framework

Bayesian inference is based on a different interpretation of probability, that probability is a measure of subjective belief.

A **prior probability distribution** is introduced for an unknown parameter, which is a probability distribution on the unknown parameter that captures your subjective belief about its possible values.

The **posterior probability distributuon** of the parameter is then calculated using Bayes theorem once data are observed. Analogs of confidence intervals and hypothesis tests can then be obtained through the posterior distribution.

An Example

Prior: $P \sim \text{Uniform}(0, 1)$

Data generating distribution: $X|P = p \sim \text{Binomial}(n, p)$

Posterior (via Bayes Theorem):

$$\Pr(P|X = x) = \frac{\Pr(X = x|P)\Pr(P)}{\Pr(X = x)}$$

Calculations

In the previous example, it is possible to analytically calculate the posterior distribution. (In the example, it is a Beta distribution with parameters that involve x.) However, this is often impossible.

Bayesian inference often involves complicated and intensive calculations to numerically approximate the posterior probability distribution.

Use in Practice

Although the Bayesian inference framework has its roots in the subjective view of probability, in modern times this philosophical aspect is often ignored or unimportant.

Instead, Bayesian inference is used because it provides a flexible and sometimes superior model for real world problems.

Extras

License

Source Code

https://github.com/SML201/lectures/tree/master/week7

Session Information

```
> sessionInfo()
R version 3.2.3 (2015-12-10)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.11.3 (El Capitan)

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
[1] ggplot2_2.1.0   knitr_1.12.3    magrittr_1.5
[4] devtools_1.10.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.3      codetools_0.2-14 digest_0.6.9
 [4] plyr_1.8.3       grid_3.2.3       gtable_0.2.0
 [7] formatR_1.2.1    evaluate_0.8     scales_0.4.0
[10] stringi_1.0-1    rmarkdown_0.9.5  labeling_0.3
[13] tools_3.2.3      stringr_1.0.0    munsell_0.4.3
[16] yaml_2.1.13      colorspace_1.2-6 memoise_1.0.0
[19] htmltools_0.3
```