I01 - Statistics

STAT 587 (Engineering)
Iowa State University

September 7, 2020

Statistics

The field of statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

https://en.wikipedia.org/wiki/Statistics

There are two different phases of statistics:

- descriptive statistics
  - statistics
  - graphical statistics
- inferential statistics
  - uses a sample to make statements about a population.

Convenience sample

The population consists of all units of interest. Any numerical characteristic of a population is a parameter. The sample consists of observed units collected from the population. Any function of a sample is called a statistic.

Population: in-use routers by graduate students at Iowa State University.

Parameter: proportion of those routers that have Gigabit speed.

Sample: students in STAT 587-2

Statistics: proportion of those students that have Gigabit routers.

Simple random sampling

A simple random sample is a sample from the population where all subsets of the same size are equally likely to be sampled. Random samples ensure that statistical conclusions will be valid.

Population: in-use routers by graduate students at Iowa State University.

Parameter: proportion of those routers that have Gigabit speed.

Sample: a pseudo-random number generator gives each graduate student a Unif(0,1) number and the lowest 100 are contacted

Statistics: proportion that have Gigabit routers.

Sampling and non-sampling errors

Sampling errors are caused by the mere fact that only a sample, a portion of a population, is observed. Fortunately,

$$\text{error} \downarrow \quad \text{as sample size } (n) \uparrow$$

Non-sampling errors are caused by inappropriate sampling schemes and wrong statistical techniques. Often, no statistical technique can rescue a poorly collected sample of data.

Sample: students in STAT 587-2

Statistics and estimators

A statistic is any function of the data.

Descriptive statistics:

- Sample mean, median, mode
- Sample quantiles
- Sample variance, standard deviation

When a statistic is meant to estimate a corresponding population parameter, we call that statistic an estimator.

Sample mean

Let $X_1, \ldots, X_n$ be a random sample from a distribution with

$$E[X_i] = \mu \qquad \text{and} \qquad Var[X_i] = \sigma^2$$

where we assume independence between the X_i.

The sample mean is

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

and estimates the population mean μ.

Sample variance

Let $X_1, \ldots, X_n$ be a random sample from a distribution with

$$E[X_i] = \mu \qquad \text{and} \qquad Var[X_i] = \sigma^2$$

where we assume independence between the X_i.

The sample variance is

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$$

and estimates the population variance σ^2.

The sample standard deviation is $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ and estimates the population standard deviation.

Quantiles

A p-quantile of a population is a number x that solves

$$P(X < x) \leq p \quad \text{and} \quad P(X > x) \leq 1 - p.$$

A sample p-quantile is any number that exceeds at most $100p\%$ of the sample, and is exceeded by at most $100(1-p)\%$ of the sample. A $100p$-percentile is a p-quantile. First, second, and third quartiles are the 25th, 50th, and 75th percentiles. They split a population or a sample into four equal parts. A median is a 0.5-quantile, 50th percentile, and 2nd quartile.

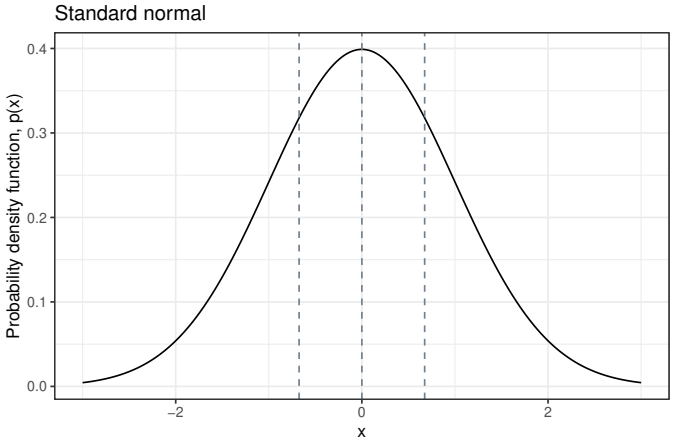
The interquartile range is the third quartile minus the first quartile, i.e.

$$IQR = Q_3 - Q_1$$

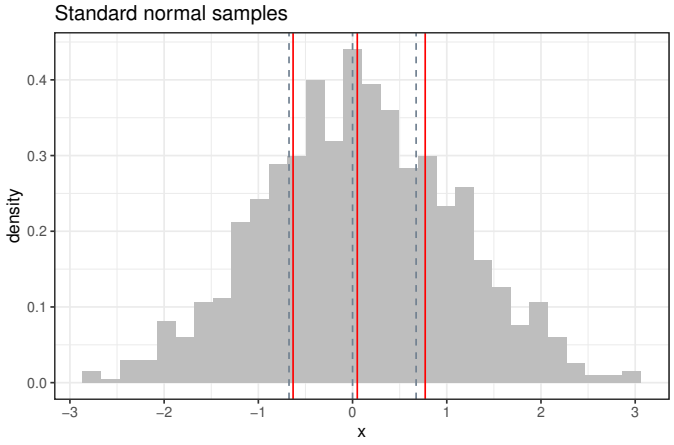
and the sample interquartile range is the third sample quartile minus the first sample quartile, i.e.

$$\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1$$

Standard normal quartiles



Sample quartiles from a standard normal



Properties of statistics and estimators

Statistics can have properties, e.g.

- standard error

Estimators can have properties, e.g.

- unbiased
- consistent

Standard error

The standard error of a statistic $\hat{\theta}$ is the standard deviation of that statistic (when the data are considered random).

If X_i are independent and have $Var[X_i] = \sigma^2$, then

$$\begin{aligned} Var\left[\overline{X}\right] &= Var\left[\tfrac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \tfrac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \tfrac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

and thus

$$SD\left[\overline{X}\right] = \sqrt{Var\left[\overline{X}\right]} = \sigma/\sqrt{n}.$$

Thus the standard error of the sample mean is $\sigma/\sqrt{n}$.

Unbiased

An estimator $\hat{\theta}$ is unbiased for a parameter θ if its expectation (when the data are considered random) equals the parameter, i.e.

$$E[\hat{\theta}] = \theta.$$

The sample mean is unbiased for the population mean μ since

$$E\left[\,\overline{X}\,\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu.$$

and the sample variance is unbiased for the population variance σ^2.

Consistent

An estimator $\hat{\theta}$, or $\hat{\theta}_n(x)$, is consistent for a parameter θ if the probability of its sampling error of any magnitude converges to 0 as the sample size n increases to infinity, i.e.

$$P\left(\left|\hat{\theta}_n(X) - \theta\right| > \epsilon\right) \to 0 \text{ as } n \to \infty$$

for any $\epsilon > 0$.

The sample mean is consistent for μ since
$Var\left[\overline{X}\right] = \sigma^2/n$ and

$$P\left(\left|\overline{X} - \mu\right| > \epsilon\right) \leq \frac{Var\left[\overline{X}\right]}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} \to 0$$

where the inequality is from Chebyshev's inequality.

Binomial example

Suppose $Y \sim Bin(n, \theta)$ where θ is the probability of success. The statistic $\hat{\theta} = Y/n$ is an estimator of θ.

Since

$$E\left[\hat{\theta}\right] = E\left[\frac{Y}{n}\right] = \frac{1}{n}E[Y] = \frac{1}{n}n\theta = \theta$$

the estimator is unbiased.

Binomial example

Suppose $Y \sim Bin(n, \theta)$ where θ is the probability of success. The statistic $\hat{\theta} = Y/n$ is an estimator of θ.

The variance of the estimator is

$$Var\left[\hat{\theta}\right] = Var\left[\frac{Y}{n}\right] = \frac{1}{n^2} Var[Y] = \frac{1}{n^2} n\theta(1-\theta) = \frac{\theta(1-\theta)}{n}.$$

Thus the standard error is

$$SE(\hat{\theta}) = \sqrt{Var[\hat{\theta}]} = \sqrt{\frac{\theta(1-\theta)}{n}}.$$

By Chebychev's inequality, this estimator is consistent for θ.

Summary

- Statistics are functions of data.
- Statistics have some properties:
  - Standard error
- Estimators are statistics that estimate population parameters.
- Estimators may have properties:
  - Unbiased
  - Consistent

Look at it!

Before you do anything with a data set, LOOK AT IT!

Why should you look at your data?

1. Find errors
   - Do variables have the correct range, e.g. positive?
   - How are Not Available encoded?
   - Are there outliers?
2. Do known or suspected relationships exist?
   - Is X linearly associated with Y?
   - Is X quadratically associated with Y?
3. Are there new relationships?
   - What is associated with Y and how?
4. Do variables adhere to distributional assumptions?
   - Does Y have an approximately normal distribution?
   - Right/left skew
   - Heavy tails

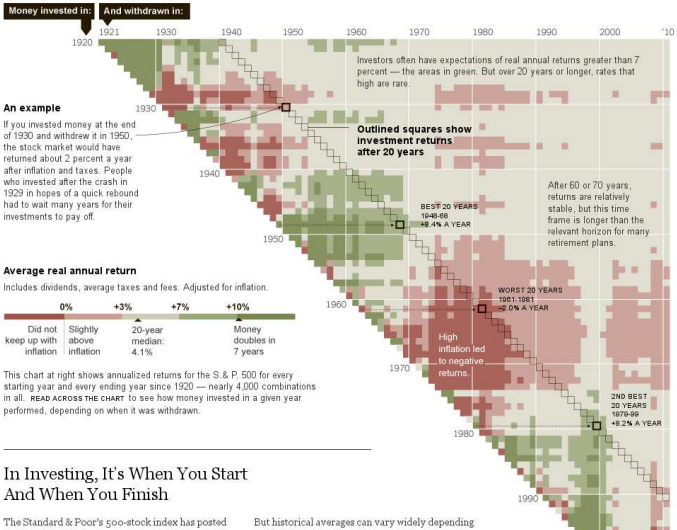
Principles of professional statistical graphics

https://moz.com/blog/data-visualization-principles-lessons-from-tufte

- Show the data
  - Avoid distorting the data, e.g. pie charts, 3d pie charts, exploding wedge 3d pie charts, bar charts that do not start at zero
- Plots should be self-explanatory
  - Use informative caption, legend
  - Use normative colors, shapes, etc
- Have a high information to ink ratio
  - Avoid bar charts
- Encourage eyes to compare
  - Use size, shape, and color to highlight differences

Stock market return

http://www.nytimes.com/interactive/2011/01/02/business/20110102-metrics-graphic.html?_r=0



In Investing, It's When You Start And When You Finish

The Standard & Poor's 500-stock index has posted

But historical averages can vary widely depending

I02 - Likelihood

STAT 587 (Engineering)
Iowa State University

September 10, 2020

Statistical modeling

A statistical model is a pair $(\mathcal{S}, \mathcal{P})$ where $\mathcal{S}$ is the set of possible observations, i.e. the sample space, and $\mathcal{P}$ is a set of probability distributions on $\mathcal{S}$.

Typically, assume a parametric model

$$p(y|\theta)$$

where

- y is our data and
- θ is unknown parameter vector.

The

- allowable values for θ determine $\mathcal{P}$ and
- the support of $p(y|\theta)$ is the set $\mathcal{S}$.

Binomial model

Suppose we will collect data were we have

- the number of success y
- out of some number of attempts n
- where each attempt is independent
- with a common probability of success θ.

Then a reasonable statistical model is

$$Y \sim Bin(n, \theta).$$

Formally,

- $\mathcal{S} = \{0, 1, 2, \ldots, n\}$ and
- $\mathcal{P} = \{Bin(n, \theta) : 0 < \theta < 1\}$.

Normal model

Suppose we have one datum

- real number,
- has a mean μ and variance σ^2, and
- uncertainty is represented by a bell-shaped curve.

Then a reasonable statistical model is

$$Y \sim N(\mu, \sigma^2).$$

Marginally,

- $\mathcal{S} = \{y : y \in \mathbb{R}\}$
- $\mathcal{P} = \{N(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$ where $\theta = (\mu, \sigma^2)$.

Normal model

Suppose our data are

- n real numbers,
- each has a mean μ and variance is σ^2,
- a histogram is reasonably approximated by a bell-shaped curve, and
- each observation is independent of the others.

Then a reasonable statistical model is

$$Y_i \overset{ind}{\sim} N(\mu, \sigma^2).$$

Marginally,

- $\mathcal{S} = \{(y_1, \ldots, y_n) : y_i \in \mathbb{R}, i \in \{1, 2, \ldots, n\}\}$
- $\mathcal{P} = \{N_n(\mu, \sigma^2 \mathrm{I}) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$ where $\theta = (\mu, \sigma^2)$.

Likelihood

The likelihood function, or simply likelihood, is the joint probability mass/density function for fixed data when viewed as a function of the parameter (vector) θ. Generically, let $p(y|\theta)$ be the joint probability mass/density function of the data and thus the likelihood is

$$L(\theta) = p(y|\theta)$$

but where y is fixed and known, i.e. it is your data.

The log-likelihood is the (natural) logarithm of the likelihood, i.e.

$$\ell(\theta) = \log L(\theta).$$

Intuition: The likelihood describes the relative support in the data for different values for your parameter, i.e. the larger the likelihood is the more consistent that parameter value is with the data.

Binomial likelihood

Suppose $Y \sim Bin(n, \theta)$, then

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

where θ is considered fixed (but often unknown) and the argument to this function is y.

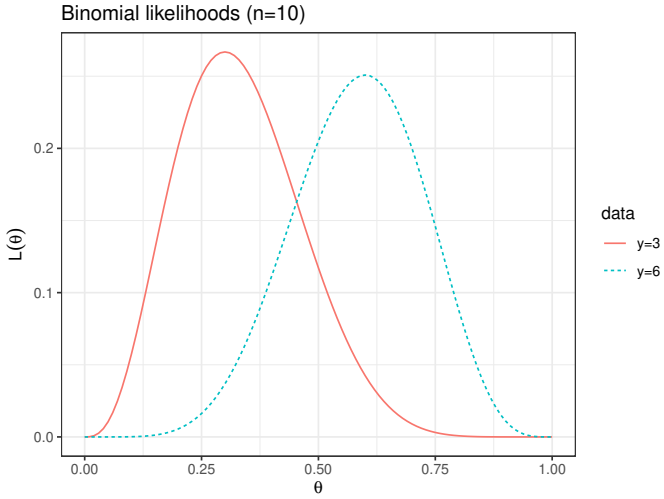
Thus the likelihood is

$$L(\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

where y is considered fixed and known and the argument to this function is θ.

Note: I write $L(\theta)$ without any conditioning, e.g. on y, so that you don't confuse this with a probability mass (or density) function.

Binomial likelihood



Likelihood for independent observations

Suppose Y_i are independent with marginal probability mass/density function $p(y_i|\theta)$.

The joint distribution for $y = (y_1, \ldots, y_n)$ is

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

The likelihood for θ is

$$L(\theta) = p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

where we are thinking about this as a function of θ for fixed y.

Normal model

Suppose $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

$$p(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

and

$$\begin{aligned} p(y|\mu, \sigma^2) &= \prod_{i=1}^n p(y_i|\mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2} \end{aligned}$$

where μ and σ^2 are fixed (but often unknown) and the argument to this function is $y = (y_1, \ldots, y_n)$.

Normal likelihood

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

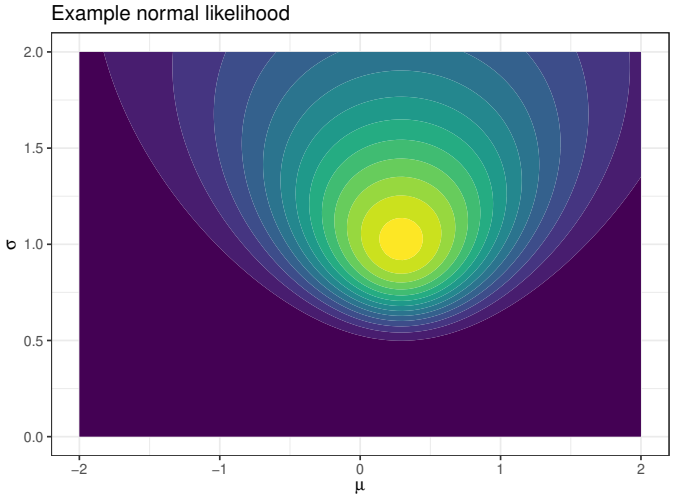
$$p(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

The likelihood is

$$L(\mu, \sigma) = p(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

where y is fixed and known and μ and σ^2 are the arguments to this function.

Normal likelihood - example contour plot



Maximum likelihood estimator (MLE)

Definition

The maximum likelihood estimator (MLE), $\hat{\theta}_{MLE}$ is the parameter value θ that maximizes the likelihood function, i.e.

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}\, L(\theta).$$

When the data are discrete, the MLE maximizes the probability of the observed data.

Binomial MLE - derivation

If $Y \sim Bin(n, \theta)$, then

$$L(\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

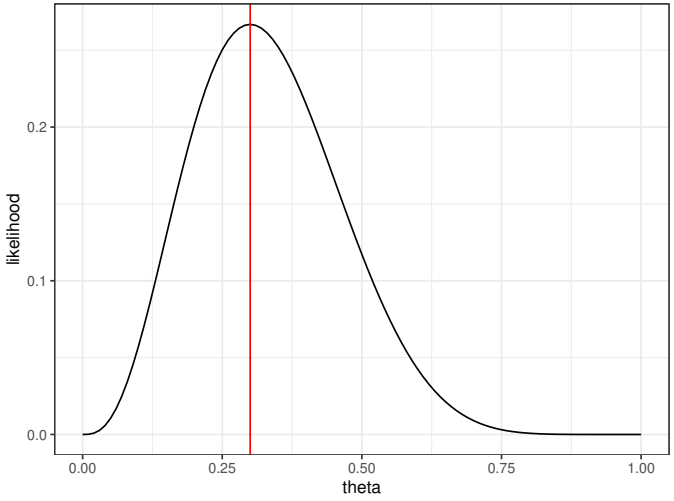
To find the MLE,

1. Take the derivative of $\ell(\theta)$ with respect to θ.
2. Set it equal to zero and solve for θ.

$$\begin{aligned} \ell(\theta) &= \log \tbinom{n}{y} + y \log(\theta) + (n-y) \log(1-\theta) \\ \tfrac{d}{d\theta}\ell(\theta) &= \tfrac{y}{\theta} - \tfrac{n-y}{1-\theta} \overset{set}{=} 0 \implies \\ \hat{\theta}_{MLE} &= y/n \end{aligned}$$

Take the second derivative of $\ell(\theta)$ with respect to θ and check to make sure it is negative.

Binomial MLE - graphically



Binomial MLE - Numerical maximization

```r
log_likelihood <- function(theta) {
  dbinom(3, size = 10, prob = theta, log = TRUE)
}

o <- optim(0.5, log_likelihood,
           method='L-BFGS-B',            # this method to use bounds
           lower = 0.001, upper = .999,  # cannot use 0 and 1 exactly
           control = list(fnscale = -1)) # maximize

o$convergence # 0 means convergence was achieved

[1] 0

o$par         # MLE

[1] 0.3000006

o$value       # value of the likelihood at the MLE

[1] -1.321151
```

Normal MLE - derivation

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

$$
\begin{aligned}
L(\mu, \sigma^2) &= \tfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2} \\
&= \tfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\overline{y}+\overline{y}-\mu)^2} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^n \left[(y_i-\overline{y})^2 + 2(y_i-\overline{y})(\overline{y}-\mu) + (\overline{y}-\mu)^2\right]\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\overline{y})^2 + -\tfrac{n}{2\sigma^2}(\overline{y}-\mu)^2\right) \quad \text{since } \textstyle\sum_{i=1}^n (y_i-\overline{y}) = 0
\end{aligned}
$$

$$
\ell(\mu, \sigma^2) = -\tfrac{n}{2}\log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\overline{y})^2 - \tfrac{1}{2\sigma^2} n(\overline{y}-\mu)^2
$$

$$
\tfrac{\partial}{\partial\mu}\ell(\mu, \sigma^2) = \tfrac{n}{\sigma^2}(\overline{y}-\mu) \overset{set}{=} 0 \implies \hat{\mu}_{MLE} = \overline{y}
$$

$$
\begin{aligned}
\tfrac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) &= -\tfrac{n}{2\sigma^2} + \tfrac{1}{2(\sigma^2)^2}\sum_{i=1}^n (y_i-\overline{y})^2 \overset{set}{=} 0 \\
&\implies \hat{\sigma}^2_{MLE} = \tfrac{1}{n}\sum_{i=1}^n (y_i-\overline{y})^2 = \tfrac{n-1}{n} S^2
\end{aligned}
$$

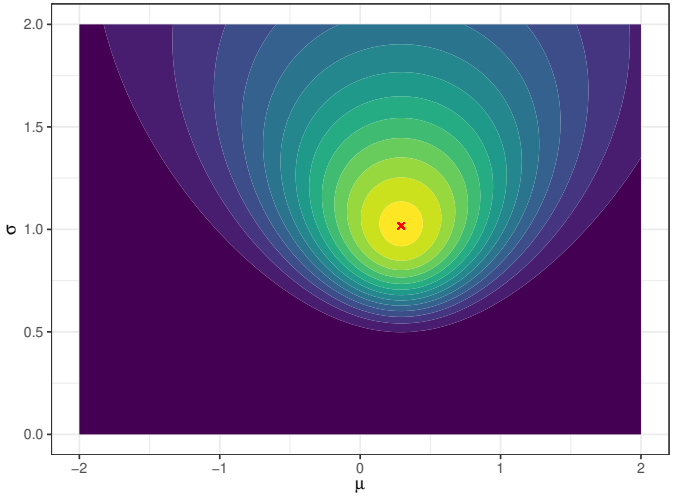
Thus, the MLE for a normal model is

$$
\hat{\mu}_{MLE} = \overline{y}, \quad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (y_i-\overline{y})^2
$$

Normal MLE - numerical maximization

```
x

[1] -0.8969145  0.1848492  1.5878453

log_likelihood <- function(theta) {
  sum(dnorm(x, mean = theta[1], sd = exp(theta[2]), log = TRUE))
}

o <- optim(c(0,0), log_likelihood,
           control = list(fnscale = -1))
c(o$par[1], exp(o$par[2])^2)                # numerical MLE

[1] 0.2918674 1.0344601

n <- length(x); c(mean(x), (n-1)/n*var(x)) # true MLE

[1] 0.2919267 1.0347381
```

Normal likelihood - graph



Summary

- For independent observations, the joint probability mass (density) function is the product of the marginal probability mass (density) functions.
- The likelihood is the joint probability mass (density) function when the argument of the function is the parameter (vector).
- The maximum likelihood estimator (MLE) is the value of the parameter (vector) that maximizes the likelihood.

I03 - Bayesian parameter estimation

STAT 587 (Engineering)
Iowa State University

September 15, 2020

Outline

- Bayesian parameter estimation
  - Condition on what is known
  - Describe **belief** using probability
  - Terminology
    - Prior $\rightarrow$ posterior
    - Posterior expectation
    - Credible intervals
  - Binomial example
    - Beta distribution

A Bayesian statistician

Let

- y be the data we will collect from an experiment,
- K be everything we know for certain about the world (aside from y), and
- θ be anything we don't know for certain.

My definition of a Bayesian statistician is an individual who makes decisions based on the probability distribution of those things we don't know conditional on what we know, i.e.

$$p(\theta|y, K).$$

Typically, the K is dropped from the notation.

Bayes' Rule

Bayes' Rule applied to a partition $P = \{A_1, A_2, \ldots\}$,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}$$

Bayes' Rule also applies to probability density (or mass) functions, e.g.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

where the integral plays the role of the sum in the previous statement.

Parameter estimation

Let y be data from some model with unknown parameter (vector) θ. Then

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

and we use the following terminology

Terminology	Notation
Posterior	$p(\theta\|y)$
Prior	$p(\theta)$
Model (likelihood)	$p(y\|\theta)$
Prior predictive (marginal likelihood)	$p(y)$

Bayesian parameter estimation involves updating your prior belief about θ, $p(\theta)$, into a posterior belief about θ, $p(\theta|y)$, based on the data observed.

Bayesian notation

We now have two distributions for our parameter θ: prior and posterior. To distinguish these two, we will have no conditioning in the prior and we will condition on y in the posterior. For example,

	Prior	Posterior
Density	$p(\theta)$	$p(\theta\|y)$
Expectation	$E[\theta]$	$E[\theta\|y]$
Variance	$Var[\theta]$	$Var[\theta\|y]$
Probabilities	$P(\theta < c)$	$P(\theta < c\|y)$

Binomial model

Suppose $Y \sim Bin(n, \theta)$, then

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

A reasonable default prior is the uniform distribution on the interval $(0, 1)$

$$p(\theta) = \mathrm{I}(0 < \theta < 1).$$

Using Bayes Rule, you can find

$$\theta|y \sim Be(1+y, 1+n-y).$$

Beta distribution

The beta distribution defines a distribution for a probability, i.e. a number on the interval (0,1). The probability density function is

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{Beta(a,b)}\mathrm{I}(0 < \theta < 1)$$

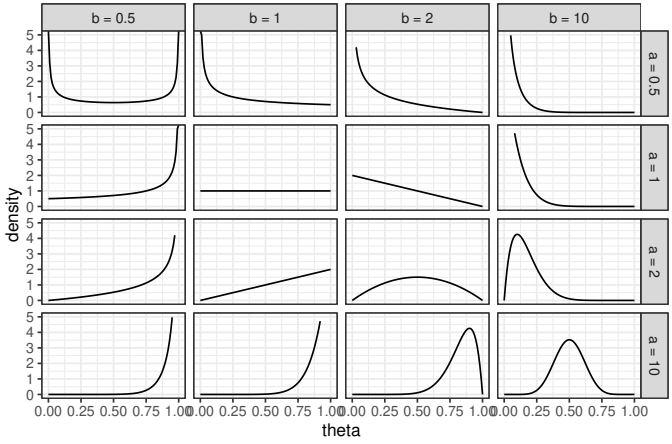
where $a, b > 0$ and $Beta$ is the beta function, i.e.

$$Beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \text{and} \quad \Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx.$$

The beta distribution has the following properties:

- $E[\theta] = \frac{a}{a+b}$,
- $Var[\theta] = \frac{ab}{(a+b)^2(a+b+1)}$, and
- $Be(1,1) \stackrel{d}{=} Unif(0,1)$.

Beta densities

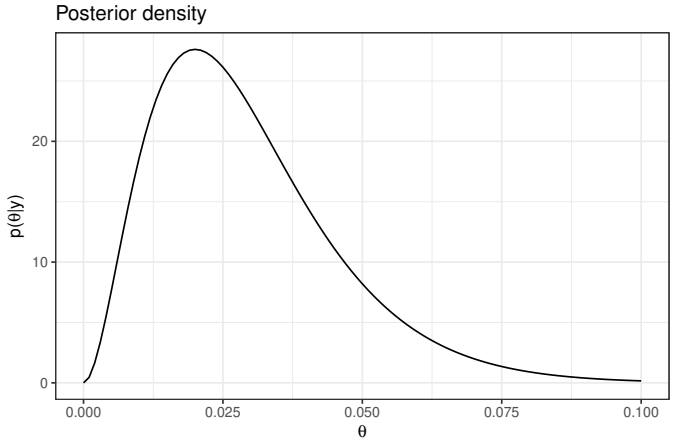


Beta posterior

Suppose we have made 100 sensors according to a particular protocol and 2 have a sensitivity below a pre-determined threshold. Let Y be the number below the threshold. Assume $Y \sim Bin(n, \theta)$ with $n = 100$ and $\theta \sim Be(1, 1)$, then

$$\theta | y \sim Be(1 + y, 1 + n - y) \stackrel{d}{=} Be(3, 99).$$

Posterior density



Posterior expectation

Often times it is inconvenient to provide a full posterior and so we often summarize using a point estimate from the posterior. For a point estimate, we can use the posterior expectation:

$$\hat{\theta}_{Bayes} = E[\theta|y] = \frac{1+y}{(1+y)+(1+n-y)} = \frac{1+y}{2+n}$$

```
(1+y)/(2+n)

[1] 0.02941176
```

Note that this is close, but not exactly equal to $\hat{\theta}_{MLE} = y/n$. Since the MLE is unbiased, this posterior expectation will generally be biased but it is still consistent since $\hat{\theta}_{Bayes} \to \hat{\theta}_{MLE}$.

Credible intervals

A $100(1-a)\%$ credible interval is any interval (L, U) such that

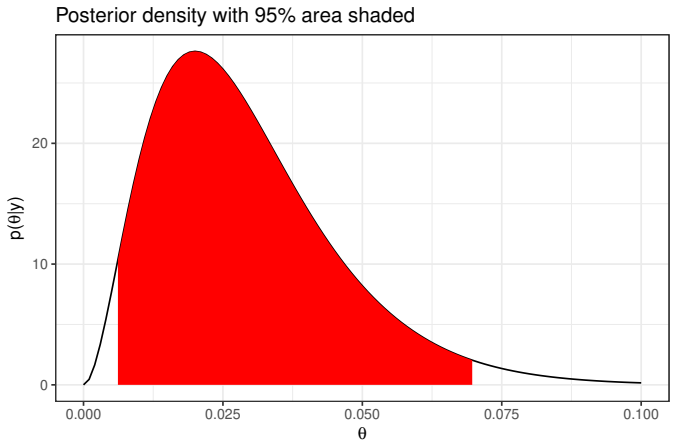
$$1 - a = \int_L^U p(\theta|y)d\theta.$$

An equal-tail $100(1-a)\%$ credible interval is the interval $L, U)$ such that

$$a/2 = \int_{-\infty}^{L} p(\theta|y)d\theta = \int_U^{\infty} p(\theta|y)d\theta.$$

```
# 95% credible interval is
ci = qbeta(c(.025,.975), 1+y, 1+n-y)
round(ci, 3)

[1] 0.006 0.070
```

Equal-tail 95% credible interval



Summary

Bayesian parameter estimation involves

1. Specifying a model $p(y|\theta)$ for your data.
2. Specifying a prior $p(\theta)$ for the parameter.
3. Deriving the posterior

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta).$$

   This equation updates your prior belief, $p(\theta)$, about the unknown parameter θ into your posterior belief, $p(\theta|y)$, about θ.
4. Calculating quantities of interest, e.g.
   - Posterior expectation, $E[\theta|y]$
   - Credible interval

Bayesian analysis for binomial model summary

Let $Y \sim Bin(n, \theta)$ and assume $\theta \sim Be(a, b)$. Then

$$\theta|y \sim Be(a + y, b + n - y).$$

A default prior is $\theta \sim Be(1, 1) \stackrel{d}{=} Unif(0, 1)$.

R code for binomial analysis:

```
a <- 1; b <- 1                    # default uniform prior
y <- 3; n <- 10                   # data

curve(dbeta(x,ay,b+n-y))          # posterior (pdf)
(a+y)/(a+b+n)                     # posterior mean
qbeta(.5, a+y, b+n-y)             # posterior median
qbeta(c(.025,.975), a+y, b+n-y) # 95% equal tail credible interval

# Probabilities
pbeta(0.5, a+y, b+n-y)            # P(theta<0.5|y)

# Special cases
qbeta(c(0,.95), a+y, b+n-y)       # if y=0, use a lower one-sided CI
qbeta(c(.05,1), a+y, b+n-y)       # if y=n, use a upper one-sided CI
```

Exponential distribution

STAT 587 (Engineering)
Iowa State University

September 17, 2020

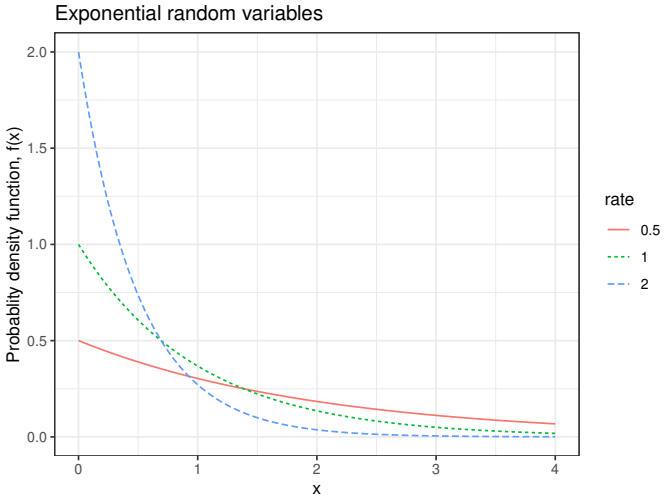
Exponential distribution

The random variable X has an exponential distribution with rate parameter $\lambda > 0$ if its probability density function is

$$p(x|\lambda) = \lambda e^{-\lambda x} \, \mathrm{I}(x > 0).$$

We write $X \sim Exp(\lambda)$.

Exponential probability density function



Exponential mean and variance

If $X \sim Exp(\lambda)$, then

$$E[X] = \int_0^\infty x\, \lambda e^{-\lambda x} dx = \cdots = \frac{1}{\lambda}$$

and

$$Var[X] = \int_0^\infty \left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x} dx = \cdots = \frac{1}{\lambda^2}.$$

Exponential cumulative distribution function

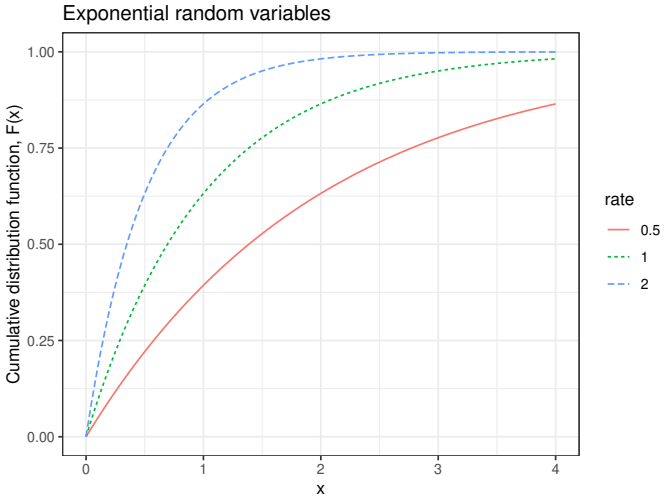
If $X \sim Exp(\lambda)$, then its cumulative distribution function is

$$F(x) = \int_0^x \lambda e^{-\lambda t} dt = \cdots = 1 - e^{-\lambda x}.$$

The inverse cumulative distribution function is

$$F^{-1}(p) = \frac{-\log(1-p)}{\lambda}.$$

Exponential cumulative distribution function - graphically



Memoryless property

Let $X \sim Exp(\lambda)$, then

$$P(X > x + c | X > c) = P(X > x).$$

Parameterization by the scale

A common alternative parameterization of the exponential distribution uses the scale $\beta = \frac{1}{\lambda}$. In this parameterization, we have

$$f(x) = \frac{1}{\beta} e^{-x/\beta} \, \mathrm{I}(x > 0)$$

and

$$E[X] = \beta \qquad \text{and} \qquad Var[X] = \beta^2.$$

Summary

Exponential random variable

- $X \sim Exp(\lambda)$, $\lambda > 0$
- $f(x) = \lambda e^{-\lambda x}$, $x > 0$
- $F(x) = 1 - e^{-\lambda x}$
- $F^{-1}(p) = \frac{-\log(1-p)}{\lambda}$
- $E[X] = \frac{1}{\lambda}$
- $Var[X] = \frac{1}{\lambda^2}$

Gamma distribution

STAT 587 (Engineering)
Iowa State University

September 17, 2020

Gamma distribution

The random variable X has a gamma distribution with

- shape parameter $\alpha > 0$ and
- rate parameter $\lambda > 0$

if its probability density function is

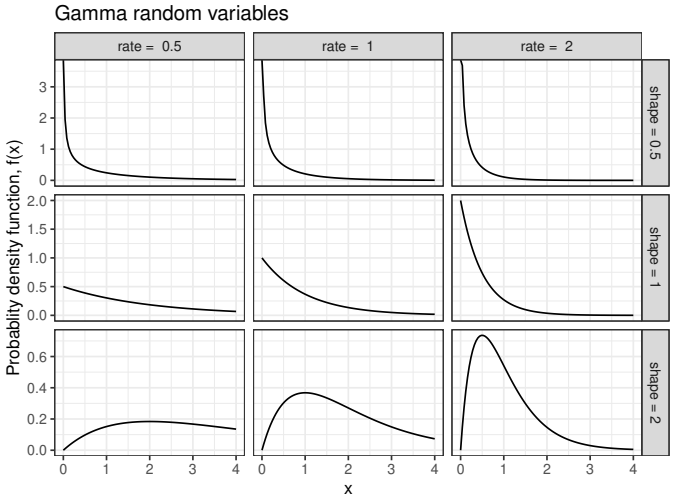
$$p(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \, \mathrm{I}(x > 0)$$

where $\Gamma(\alpha)$ is the gamma function,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

We write $X \sim Ga(\alpha, \lambda)$.

Gamma probability density function



Gamma mean and variance

If $X \sim Ga(\alpha, \lambda)$, then

$$E[X] = \int_0^\infty x \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx = \cdots = \frac{\alpha}{\lambda}$$

and

$$Var[X] = \int_0^\infty \left(x - \frac{\alpha}{\lambda}\right)^2 \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx = \cdots = \frac{\alpha}{\lambda^2}.$$

Gamma cumulative distribution function

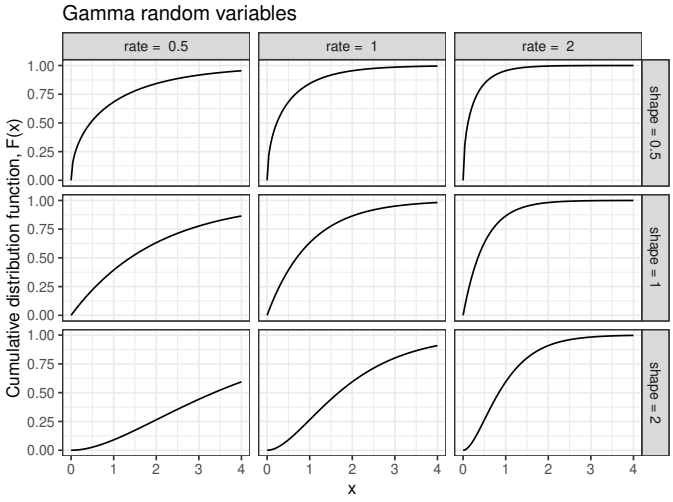
If $X \sim Ga(\alpha, \lambda)$, then its cumulative distribution function is

$$F(x) = \int_0^x \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t} dt = \cdots = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}$$

where $\gamma(\alpha, \beta x)$ is the incomplete gamma function, i.e.

$$\gamma(\alpha, \beta x) = \int_0^{\beta x} t^{\alpha-1} e^{-t} dt.$$

Gamma cumulative distribution function - graphically



Relationship to exponential distribution

If $X_i \overset{iid}{\sim} Exp(\lambda)$, then

$$Y = \sum_{i=1}^{n} X_i \sim Ga(n, \lambda).$$

Thus, $Ga(1, \lambda) \overset{d}{=} Exp(\lambda)$.

Parameterization by the scale

A common alternative parameterization of the Gamma distribution uses the scale $\theta = \frac{1}{\lambda}$. In this parameterization, we have

$$f(x) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta} \, \mathrm{I}(x > 0)$$

and

$$E[X] = \alpha\theta \qquad \text{and} \qquad Var[X] = \alpha\theta^2.$$

Summary

Gamma random variable

- $X \sim Ga(\alpha, \lambda)$, $\alpha, \lambda > 0$
- $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $x > 0$
- $E[X] = \frac{\alpha}{\lambda}$
- $Var[X] = \frac{\alpha}{\lambda^2}$

Inverse gamma distribution

STAT 587 (Engineering)
Iowa State University

September 17, 2020

Inverse gamma distribution

The random variable X has an inverse gamma distribution with

- shape parameter $\alpha > 0$ and
- scale parameter $\beta > 0$

if its probability density function is

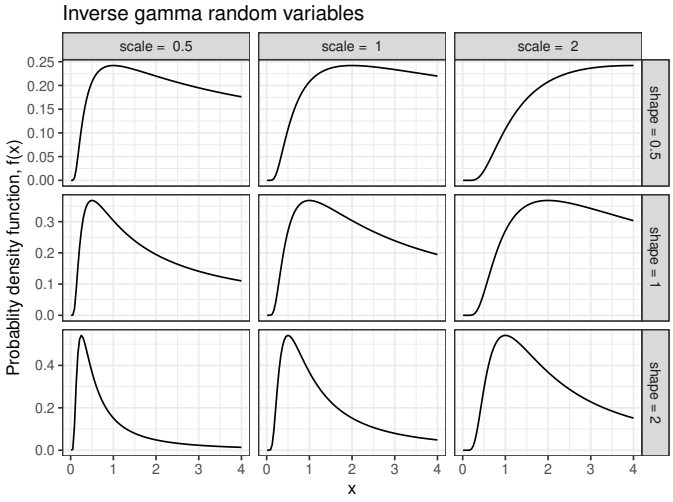
$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x} \, \mathrm{I}(x > 0).$$

where $\Gamma(\alpha)$ is the gamma function,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

We write $X \sim IG(\alpha, \beta)$.

Inverse gamma probability density function



Inverse gamma mean and variance

If $X \sim IG(\alpha, \beta)$, then

$$E[X] = \int_0^\infty x \, \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x} dx = \cdots = \frac{\beta}{\alpha - 1}, \quad \alpha > 1$$

and

$$\begin{aligned} Var[X] &= \int_0^\infty \left(x - \tfrac{\beta}{\alpha-1} \right)^2 \tfrac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x} dx \\ &= \cdots = \tfrac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2. \end{aligned}$$

Relationship to gamma distribution

If $X \sim Ga(\alpha, \lambda)$ where λ is the rate parameter, then

$$Y = \frac{1}{X} \sim IG(\alpha, \lambda).$$

Summary

Inverse gamma random variable

- $X \sim IG(\alpha, \beta)$, $\alpha, \beta > 0$
- $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}, x > 0$
- $E[X] = \frac{\beta}{\alpha-1}$, $\alpha > 1$
- $Var[X] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, $\alpha > 2$

Student's t-distribution

STAT 587 (Engineering)
Iowa State University

September 17, 2020

Student's t distribution

The random variable X has a Student's t distribution with degrees of freedom $\nu > 0$ if its probability density function is

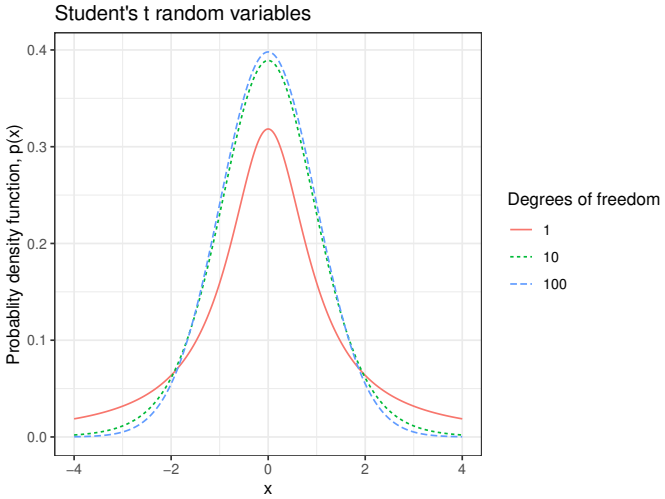
$$p(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $\Gamma(\alpha)$ is the gamma function,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx.$$

We write $X \sim t_\nu$.

Student's t probability density function



Student's t mean and variance

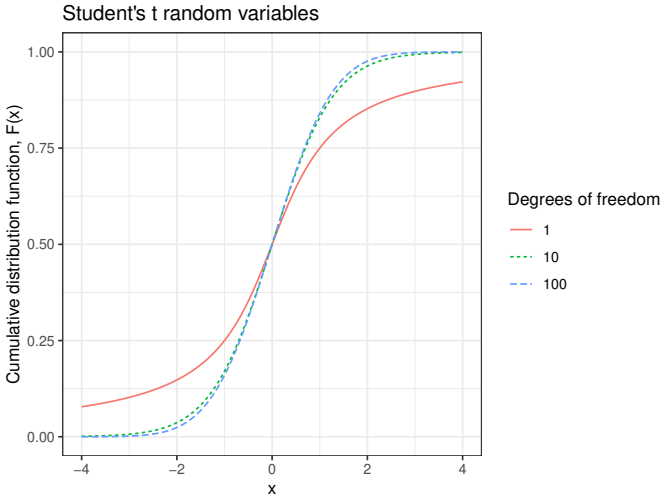
If $T \sim t_v$, then

$$E[X] = \int_{-\infty}^{\infty} x \, \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx = \cdots = 0, \quad \nu > 1$$

and

$$Var[X] = \int_{0}^{\infty} (x-0)^2 \, \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx = \cdots = \frac{\nu}{\nu-2}, \quad \nu > 2.$$

Gamma cumulative distribution function - graphically



Location-scale t distribution

If $X \sim t_\nu$, then

$$Y = \mu + \sigma X \sim t_\nu(\mu, \sigma^2)$$

for parameters:

- degrees of freedom $\nu > 0$,
- location μ and
- scale $\sigma > 0$.

By properties of expectations and variances, we can find that

$$E[Y] = \mu, \qquad \nu > 1$$

and

$$Var[Y] = \frac{\nu}{\nu - 2}\sigma^2, \qquad \nu > 2.$$

Generalized Student's t probability density function

The random variable Y has a generalized Student's t distribution with

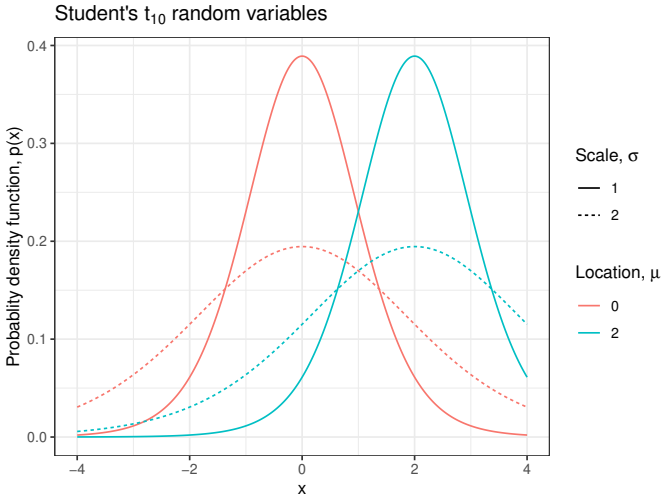
- degrees of freedom $\nu > 0$,
- location μ, and
- scale $\sigma > 0$

if its probability density function is

$$p(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma}\left(1 + \frac{1}{\nu}\left[\frac{y-\mu}{\sigma}\right]^2\right)^{-\frac{\nu+1}{2}}$$

We write $Y \sim t_\nu(\mu, \sigma^2)$.

Generalized Student's t probability density function



t with 1 degree of freedom

If $T \sim t_1(\mu, \sigma^2)$, then T has a Cauchy distribution and we write

$$T \sim Ca(\mu, \sigma^2).$$

If $T \sim t_1(0, 1)$, then T has a standard Cauchy distribution. A Cauchy random variable has no mean or variance.

As degrees of freedom increases

If $T_\nu \sim t_\nu(\mu, \sigma^2)$, then

$$\lim_{\nu \to \infty} T_\nu \overset{d}{=} X \sim N(\mu, \sigma^2)$$

t distribution arising from a normal sample

Let $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$. We calculate the sample mean

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

and the sample variance

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

Then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Inverse-gamma scale mixture of a normal

If

$$X|\sigma^2 \sim N(\mu, \sigma^2/n) \qquad \text{and} \qquad \sigma^2 \sim IG\left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right)$$

then

$$X \sim t_\nu(\mu, s^2/n)$$

which is obtained by

$$p_x(x) = \int p_{x|\sigma^2}(x|\sigma^2)p_{\sigma^2}(\sigma^2)d\sigma^2$$

where

- p_x is the marginal density for x
- $p_{x|\sigma^2}$ is the conditional density for x given σ^2, and
- p_{σ^2} is the marginal density for σ^2.

Summary

Student's t random variable:

- $T \sim t_\nu(\mu, \sigma^2)$, $\nu, \sigma > 0$
- $E[X] = \mu$, $\nu > 1$
- $Var[X] = \frac{\nu}{\nu-2}\sigma^2$, $\nu > 2$
- Relationships to other distributions

I4 - Bayesian parameter estimation in a normal model

STAT 587 (Engineering)
Iowa State University

September 18, 2020

Bayesian parameter estimation

Recall that Bayesian parameter estimation involves

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta)}$$

with

- posterior $p(\theta|y)$,
- prior $p(\theta)$,
- model $p(y|\theta)$, and
- prior predictive $p(y)$.

For this video, $\theta = (\mu, \sigma^2)$ and

$$y|\mu, \sigma^2 \sim N(\mu, \sigma^2).$$

Bayesian parameter estimation in a normal model

Let $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ and the default prior

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Note: This "prior" is not a distribution since its integral is not finite. Nonetheless, we can still derive the following posterior

$$\mu|y \sim t_{n-1}(\overline{y}, s^2/n) \qquad \text{and} \qquad \sigma^2|y \sim IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

where

- n is the sample size,
- $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean, and
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ is the sample variance.

Posterior for the mean

The posterior for the mean is

$$\mu|y \sim t_{n-1}(\overline{y}, s^2/n)$$

and from properties of the generalized Student's t distribution, we know

- $E[\mu|y] = \overline{y}$ for $n > 2$,
- $Var[\mu|y] = \frac{(n-1)s^2}{(n-3)} \Big/ n$ for $n > 3$,

and

$$\frac{\mu - \overline{y}}{s/\sqrt{n}} \sim t_{n-1}.$$

Credible intervals for μ

Since

$$\frac{\mu - \overline{y}}{s/\sqrt{n}} \sim t_{n-1}$$

a $100(1-a)\%$ equal-tail credible interval is

$$\overline{y} \pm t_{n-1,a/2}\, s/\sqrt{n}$$

where $t_{n-1,a/2}$ is a t critical value such that $P(T_{n-1} < t_{n-1,a/2}) = 1 - a/2$ when $T_{n-1} \sim t_{n-1}$.

For example, $t_{10-1,0.05/2}$ is

```
n = 10
a = 0.05 # 95\% CI
qt(1-a/2, df = n-1)
```

```
[1] 2.262157
```

Posterior for the variance

The posterior for the mean is

$$\sigma^2|y \sim IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

and from properties of the inverse Gamma distribution, we know

- $E[\sigma^2|y] = \frac{(n-1)s^2}{n-3}$ for $n > 3$,

and

$$\frac{1}{\sigma^2}\bigg| y \sim Ga\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

where $(n-1)s^2/2$ is the rate parameter.

Credible intervals for σ^2

For a $100(1-a)\%$ credible interval, we need

$$a/2 = P(\sigma^2 < L|y) = P(\sigma^2 > U|y).$$

To do this, we will find

$$a/2 = P\left(\frac{1}{\sigma^2} > \frac{1}{L}\middle| y\right) = P\left(\frac{1}{\sigma^2} < \frac{1}{U}\middle| y\right).$$

Here is a function that performs this computation

```
qinvgamma <- function(p, shape, scale = 1)
  1/qgamma(1-p, shape = shape, rate = scale)
```

Posterior for the standard deviation, σ

The variance is hard to interpret because its units are squared relative to Y_i. In contrast, the standard deviation $\sigma = \sqrt{\sigma^2}$ units are the same as Y_i.

For credible intervals (or any quantile), we can compute the square root of the endpoints since

$$P(\sigma^2 < c^2) = P(\sigma < c).$$

Find the pdf through transformations of random variables. In R code,

```
dinvgamma <- function(x, shape, scale = 1)
  dgamma(1/x, shape = shape, rate = scale)/x^2

dsqrtinvgamma = function(x, shape, scale)
  dinvgamma(x^2, shape, scale)*2*x
```

Yield data

Suppose we have a random sample of 9 Iowa farms and we obtain corn yield in bushels per acre on those farms. Let Y_i be the yield for farm i in bushels/acre and assume

$$Y_i \overset{ind}{\sim} N(\mu, \sigma^2).$$

We are interested in making statements about μ and σ^2.

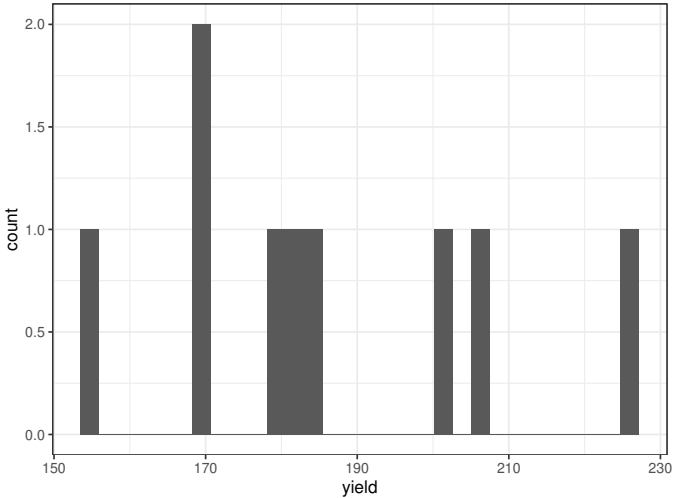
```
yield_data <- read.csv("yield.csv")
nrow(yield_data)
```

```
[1] 9
```

```
yield_data
```

```
   farm    yield
1 farm1 153.5451
2 farm2 205.6999
3 farm3 178.7548
4 farm4 170.1692
5 farm5 224.7723
6 farm6 184.0806
7 farm7 169.8615
8 farm8 201.2721
9 farm9 181.6356
```

Histogram of yield



Calculate sufficient statistics

```
n               = length(yield_data$yield); n
```

```
[1] 9
```

```
sample_mean     = mean(yield_data$yield);   sample_mean
```

```
[1] 185.5323
```

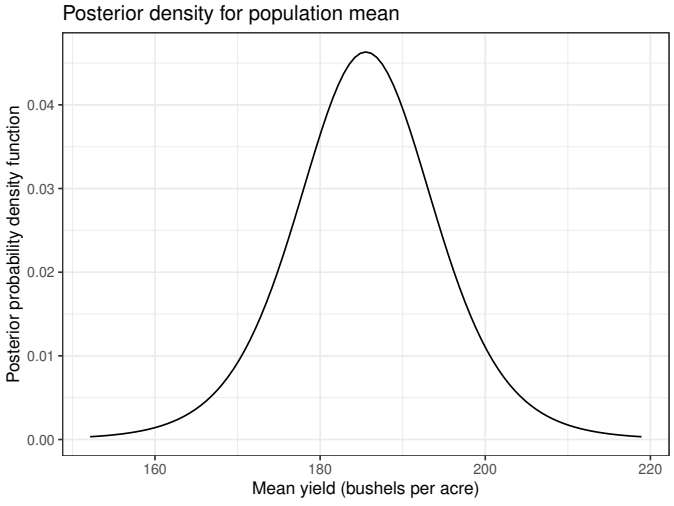
```
sample_variance = var(yield_data$yield);    sample_variance
```

```
[1] 470.2817
```

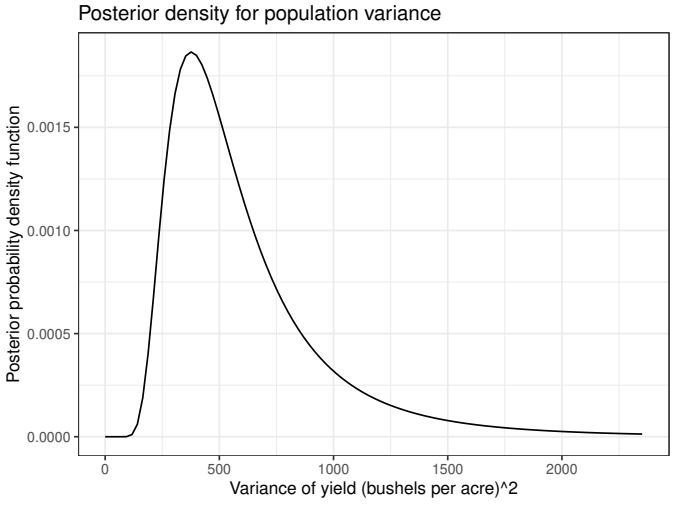
Use these sufficient statistics to calculate:

- posterior densities
- posterior means
- credible intervals

Posterior density for μ



Posterior density for σ^2



Posterior means

```
# Posterior mean of population yield mean, E[mu|y]
sample_mean
```

```
[1] 185.5323
```

Posterior mean for μ is $E[\mu|y] = 186$ bushels/acre.

```
# Posterior mean of population yield variance
post_mean_var = (n-1)*sample_variance / (n-3)
post_mean_var
```

```
[1] 627.0422
```

Posterior mean for σ^2 is $E[\sigma^2|y] = 627$ (bushels/acre)^2.

Credible intervals

```
# 95% credible interval for the population mean
a = 0.05
mean_ci = sample_mean + c(-1,1) *  qt(1-a/2, df = n-1) * sqrt(sample_variance/n)
mean_ci
```

```
[1] 168.8630 202.2017
```

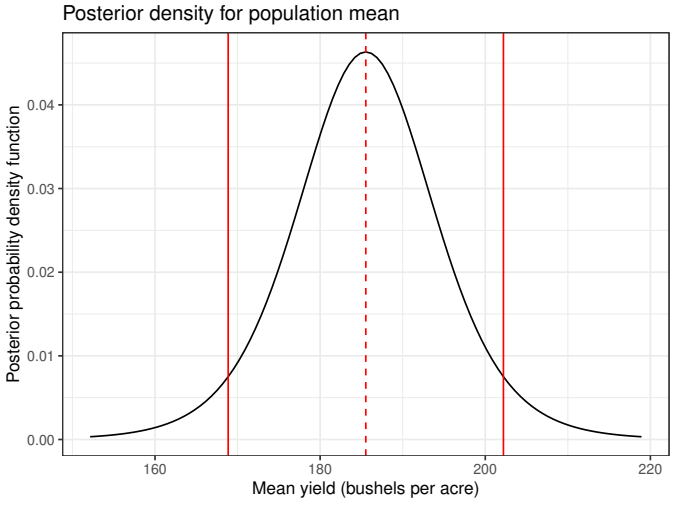
So a 95% credible interval for μ is (169,202) bushels/acre.

```
# 95% credible interval for the population variance
var_ci = qinvgamma(c(a/2, 1-a/2),
                   shape = (n-1)/2,
                   scale = (n-1)*sample_variance/2)
var_ci
```

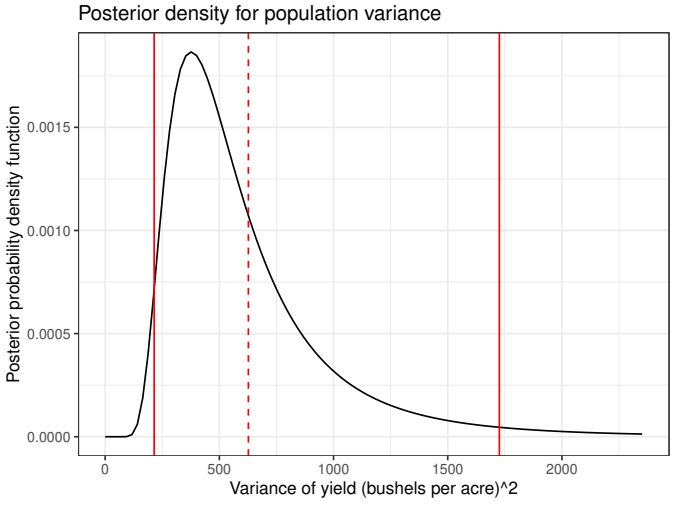
```
[1]  214.5623 1726.0175
```

So a 95% credible interval for σ^2 is (215,1726) (bushels/acre)^2

Posterior density for μ



Posterior density for σ^2



Posterior for the standard deviation, σ

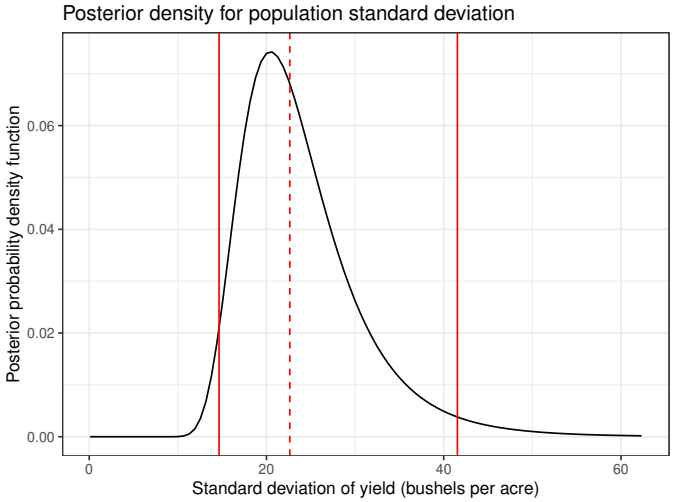
```
# Posterior median and 95% CI for population yield standard deviation
sd_median = sqrt(qinvgamma(.5, shape = (n-1)/2, scale = (n-1)*sample_variance/2))
sd_median

[1] 22.63362
```

So the posterior median for σ is 23 bushels/acre.

```
# Posterior 95% CI for the population yield standard deviation
sd_ci = sqrt(var_ci)
sd_ci

[1] 14.64795 41.54537
```

So a posterior 95% credible interval for σ is 15, 42 bushels/acre.

Posterior for the standard deviation, σ



Bayesian inference in a normal model

- Prior: $p(\mu, \sigma^2) = 1/\sigma^2$
- Posterior:

$$\mu|y \sim t_{n-1}(\overline{y}, s^2/n) \qquad \text{and} \qquad \sigma^2|y \sim IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

```
# Sufficient statistics
n               = length(y)
sample_mean     = mean(y)
sample_variance = var(y)

# Posterior expectations
sample_mean                 # mu
(n-1)*sample_variance / (n-3) # sigma^2

# Posterior medians
var_median = qinvgamma(.5, shape = (n-1)/2, scale = (n-1)*sample_variance/2)
sd_median  = sqrt(median_var)

# Posterior credible intervals
sample_mean + c(-1,1) *  qt(1-a/2, df = n-1) * sqrt(sample_variance/n)
var_ci = qinvgamma(c(a/2,1-a/2), shape = (n-1)/2, scale = (n-1)*sample_variance/2)
sd_ci  = sqrt(var_ci)
```

I05 - Confidence intervals

STAT 587 (Engineering)
Iowa State University

September 24, 2020

Exact confidence intervals

The coverage of an interval estimator is the probability the interval will contain the true value of the parameter *when the data are considered to be random*. If an interval estimator has $100(1-a)\%$ coverage, then we call it a $100(1-a)\%$ confidence interval and $1-a$ is the confidence level.

That is, we calculate

$$1 - a = P(L < \theta < U)$$

where L and U are random because they depend on the data. Thus confidence is a statement about the procedure.

Normal model

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ and we assume the default prior $p(\mu, \sigma^2) \propto 1/\sigma^2$, then a $100(1-a)\%$ credible interval for μ is given by

$$\overline{y} \pm t_{n-1,a/2} s/\sqrt{n}.$$

When the data are considered random

$$T_{n-1} = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}(0,1)$$

thus the probability μ is within our credible interval is

$$\begin{aligned} & P\left(\overline{Y} - t_{n-1,a/2} S/\sqrt{n} < \mu < \overline{Y} + t_{n-1,a/2} S/\sqrt{n}\right) \\ &= P\left(-t_{n-1,a/2} < \tfrac{\overline{Y}-\mu}{S/\sqrt{n}} < t_{n-1,a/2}\right) \\ &= P\left(-t_{n-1,a/2} < T_{n-1} < t_{n-1,a/2}\right) \\ &= 1 - a. \end{aligned}$$

Thus, this $100(1-a)\%$ credible interval is also a $100(1-a)\%$ confidence interval.

Yield data example

Recall the corn yield example from I04 with 9 randomly selected fields in Iowa whose sample average yield is 186 and sample standard deviation is 22. Then a 95% confidence interval for the mean corn yield on Iowa farms is

$$186 \pm 2.31 \times 22/\sqrt{9} = (169, 202).$$

Standard error

The standard error of an estimator is an *estimate* of the standard deviation of the estimator (when the data are considered random).

If $Y \sim Bin(n, \theta)$, then

$$\hat{\theta} = \frac{Y}{n} \qquad \text{has} \qquad SE[\hat{\theta}] = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

$$\hat{\mu} = \overline{Y} \qquad \text{has} \qquad SE[\hat{\mu}] = S/\sqrt{n}.$$

Approximate confidence intervals

If an unbiased estimator has an asymptotic normal distribution, then we can construct an approximate $100(1-a)\%$ confidence interval for $E[\hat{\theta}] = \theta$ using

$$\hat{\theta} \pm z_{a/2} SE[\hat{\theta}].$$

where $SE[\hat{\theta}]$ is the standard error of the estimator and $P(Z > z_{a/2}) = a/2$.

This comes from the fact that if $\hat{\theta} \dot{\sim} N(\theta, SE[\hat{\theta}]^2)$, then

$$\begin{aligned}
&P\left(\hat{\theta} - z_{a/2} SE(\hat{\theta}) < \theta < \hat{\theta} + z_{a/2} SE(\hat{\theta})\right) \\
&= P\left(-z_{a/2} < \tfrac{\hat{\theta}-\theta}{SE(\hat{\theta})} < z_{a/2}\right) \\
&\approx P\left(-z_{a/2} < Z < z_{a/2}\right) \\
&= 1 - a.
\end{aligned}$$

Normal example

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ and we have the estimator $\hat{\mu} = \overline{Y}$, then

$$E[\hat{\mu}] = \mu \qquad \text{and} \qquad SE[\hat{\mu}] = S/\sqrt{n}$$

Thus an approximate $100(1-a)\%$ confidence interval for $\mu = E[\hat{\mu}]$ is

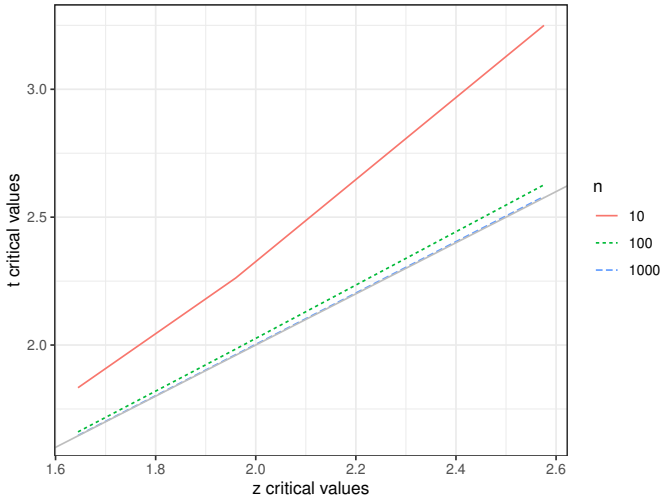
$$\hat{\mu} \pm z_{a/2} SE[\hat{\mu}] = \overline{Y} \pm z_{a/2} S/\sqrt{n}.$$

Note that this is almost identical to the exact $100(1-a)\%$ confidence interval for μ,

$$\overline{Y} \pm t_{n-1,a/2} S/\sqrt{n}$$

and when n is large $z_{a/2} \approx t_{n-1,a/2}$.

T critical values vs Z critical values



Approximate confidence interval for binomial proportion

If $Y \sim Bin(n, \theta)$, then an approximate $100(1-a)\%$ confidence interval for θ is

$$\hat{\theta} \pm z_{a/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

where $\hat{\theta} = Y/n$ since

$$E[\hat{\theta}] = E\left[\frac{Y}{n}\right] = \theta$$

and

$$SE[\hat{\theta}] = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

Gallup poll example

In a Gallup poll dated 2017/02/19, 32.1% of respondents of the 1,500 randomly selected U.S. adults indicated that they were "engaged at work". Thus an approximate 95% confidence interval for the proportion of all U.S. adults is

$$0.321 \pm 1.96 \times \sqrt{\frac{.321(1-.321)}{1500}} = (0.30, 0.34).$$

Confidence interval summary

Model	Parameter	Estimator	Confidence Interval	Type
$Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$	μ	$\hat{\mu} = \overline{y}$	$\hat{\mu} \pm t_{n-1,a/2} s/\sqrt{n}$	exact
$Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$	μ	$\hat{\mu} = \overline{y}$	$\hat{\mu} \pm z_{a/2} s/\sqrt{n}$	approximate
$Y \sim Bin(n, \theta)$	θ	$\hat{\theta} = y/n$	$\hat{\theta} \pm z_{a/2}\sqrt{\hat{\theta}(1-\hat{\theta})/n}$	approximate
$Y_i \overset{ind}{\sim} Ber(\theta)$	θ	$\hat{\theta} = \overline{y}$	$\hat{\theta} \pm z_{a/2}\sqrt{\hat{\theta}(1-\hat{\theta})/n}$	approximate

Bayesian credible intervals generally provide approximate confidence intervals.

Approximate means that the coverage will get closer to the desired probability, i.e. $100(1-a)\%$, as the sample size gets larger.

I06 - p-values

STAT 587 (Engineering)
Iowa State University

September 27, 2020

p-value

A p-value is the probability of observing a statistic as or more extreme than observed if the model is true.

A p-value is the probability of observing a statistic as or more extreme than *the one you* observed if the model is true *when the data are considered random*.

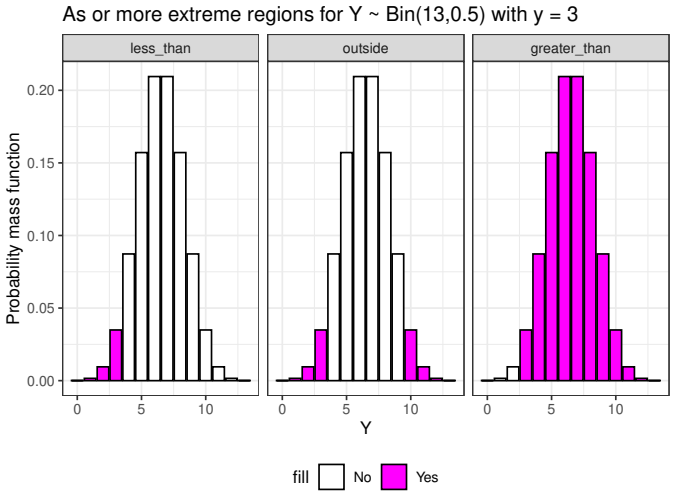
Binomial model

Let $H_0 : Y \sim Bin(13, 0.5)$ and observe $y = 3$.

Choos

- statistic is 3,
- its sampling distribution *when the model is true is* $Y \sim Bin(13, 0.5)$, and
- there are three *as or more extreme regions*:
  - $Y \leq 3$
  - $Y \geq 3$
  - $|Y - 13 \cdot 0.5| \geq |3 - 13 \cdot 0.5|$

as or more extreme regions



R Calculation

One-sided p-values:

- $P(Y \leq y)$:

```
pbinom(y, size = n, prob = p)
[1] 0.04614258
```

- $P(Y \geq y) = 1 - P(Y < y) = 1 - P(Y \leq y - 1)$:

```
1-pbinom(y-1, size = n, prob = p)
[1] 0.9887695
```

Two-sided p-value:

$$P(|Y - n\theta| \leq |y - n\theta|) = 2P(Y \leq y)$$

```
2*pbinom(y, size = n, prob = p)
[1] 0.09228516
```

Normal model

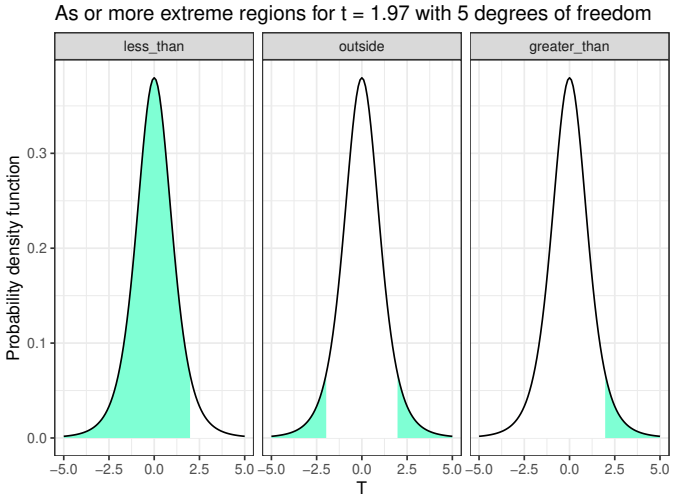
Let $H_0 : Y_i \sim N(3, 4^2)$ for $i = 1, \ldots, 6$ and you observe $\overline{y} = 6.3$, $s = 4.1$, and

$$t = \frac{\overline{y} - 3}{s/\sqrt{n}} = \frac{6.3 - 3}{4.1/\sqrt{6}} = 1.97.$$

Choose

- t-statistic $t = 1.97$,
- its sampling distribution *when the model is true is* $T_5 \sim t_5$, and
- there are three *as or more extreme regions*:
  - $T_5 \leq 1.97$
  - $T_5 \geq 1.97$
  - $|T_5| \geq |1.97|$

as or more extreme regions



R Calculation

- One-sided p-values:
  - $P(T_5 \leq t)$:

```
pt(t, df = n-1)
[1] 0.9471422
```

  - $P(T_5 \geq t) = 1 - P(T_5 < t) = 1 - P(T_5 \leq t)$:

```
1-pt(t, df = n-1)
[1] 0.05285775
```

- Two-sided p-value:

$$P(|T_5| \geq |t|) = 2P(T_5 \geq t)$$

```
2*(1-pt(t, df = n-1))
[1] 0.1057155
```

Interpretation

Small p-values provide evidence that the data are incompatible with the model.

Recall

$$Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$$

indicates the data

- are independent,
- are normally distributed,
- have a common mean, and
- have a common variance.

Summary

- p-value: the probability of observing a statistic as or more extreme than observed if the model is true
- small p-values provide evidence that the data are incompatible with the model

Hypothesis tests

with binomial example

STAT 587 (Engineering)
Iowa State University

October 2, 2020

Statistical hypothesis testing

A hypothesis test consists of two hypotheses,

- null hypothesis (H_0) and
- an alternative hypothesis (H_A),

which make claims about parameter(s) in a model, and a decision to either

- reject the null hypothesis or
- fail to reject the null hypothesis.

Binomial model

If $Y \sim Bin(n, \theta)$, then some hypothesis tests are

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_A : \theta \neq \theta_0$$

or

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_A : \theta > \theta_0$$

or

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_A : \theta < \theta_0$$

Small data

Let $Y \sim Bin(n, \theta)$ with

$$H_0 : \theta = 0.5 \qquad \text{versus} \qquad H_A : \theta \neq 0.5.$$

You collect data and observe $y = 6$ out of $n = 13$ attempts. Should you reject H_0? Probably not since $6 \approx E[Y] = 6.5$ if H_0 is true.

What if you observed $y = 2$? Well, $P(Y = 2) \approx 0.01$.

Large data

Let $Y \sim Bin(n, \theta)$ with

$$H_0 : \theta = 0.5 \qquad \text{versus} \qquad H_A : \theta \neq 0.5.$$

You collect data and observe $y = 6500$ out of $n = 13000$ attempts. Should you reject H_0? Probably not since $6500 = E[Y]$ if H_0 is true. But $P(Y = 6500) \approx 0.007$.

p-values

p-value: the probability of observing a test statistic as or more extreme than observed if the null hypothesis is true

The as or more extreme region is determined by the alternative hypothesis.

For example, if $Y \sim Bin(n, \theta)$ and $H_0 : \theta = \theta_0$ then

$$H_A : \theta < \theta_0 \implies Y \leq y$$

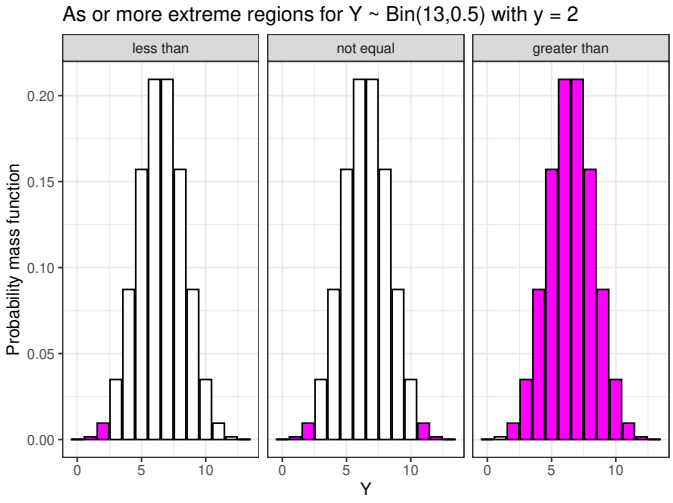
or

$$H_A : \theta > \theta_0 \implies Y \geq y$$

or

$$H_A : \theta \neq \theta_0 \implies |Y - n\theta_0| \geq |y - n\theta_0|.$$

as or more extreme regions



R "hand" calculation

$$H_A : \theta < 0.5 \implies p\text{-value} = P(Y \leq y)$$

```
pbinom(y, size = n, prob = theta0)
```

```
[1] 0.01123047
```

$$H_A : \theta > 0.5 \implies p\text{-value} = P(Y \geq y) = 1 - P(Y \leq y - 1)$$

```
1-pbinom(y-1, size = n, prob = theta0)
```

```
[1] 0.998291
```

$$H_A : \theta \neq 0.5 \implies p\text{-value} = P(|Y - n\theta_0| \leq |y - n\theta_0|)$$

```
2*pbinom(y, size = n, prob = theta0)
```

```
[1] 0.02246094
```

R Calculation

$$H_A : \theta < 0.5$$

```
binom.test(y, n, p = theta0, alternative = "less")$p.value
```

```
[1] 0.01123047
```

$$H_A : \theta > 0.5$$

```
binom.test(y, n, p = theta0, alternative = "greater")$p.value
```

```
[1] 0.998291
```

$$H_A : \theta \neq 0.5$$

```
binom.test(y, n, p = theta0, alternative = "two.sided")$p.value
```

```
[1] 0.02246094
```

Significance level

Make a decision to either

- reject the null hypothesis or
- fail to reject the null hypothesis.

Select a significance level a and

- reject if p-value $< a$ otherwise
- fail to reject.

Decisions

	Truth	
Decision	H_0 true	H_0 not true
reject H_0	type I error	correct
fail to reject H_0	correct	type II error

Then

significance level a is $P(\text{reject } H_0 | H_0 \text{ true})$

and

power is $P(\text{reject } H_0 | H_0 \text{ not true})$.

Interpretation

The null hypothesis is a model. For example,

$$H_0 : Y \sim Bin(n, \theta_0)$$

if we reject H_0, then we are saying the data are incompatible with this model.

Recall that $Y = \sum_{i=1}^{n} X_i$ for $X_i \overset{ind}{\sim} Ber(\theta)$.

So, possibly

- the X_i are not independent or
- they don't have a common θ or
- $\theta \neq \theta_0$ or
- you just got unlucky.

If we fail to reject H_0, insufficient evidence to say that the data are incompatible with this model.

Die tossing example

You are playing a game of Dragonwood and a friend rolled a four 3 times in 6 attempts. Did your friend (somehow) increase the probability of rolling a 4?

Let Y be the number of fours rolled and assume $Y \sim Bin(6, \theta)$. You observed $y = 3$ and are testing

$$H_0 : \theta = \frac{1}{6} \qquad \text{versus} \qquad H_A : \theta > \frac{1}{6}.$$

```
binom.test(3, 6, p = 1/6, alternative = "greater")$p.value

[1] 0.06228567
```

With a signficance level of $a = 0.05$, you fail to reject the null hypothesis.

Summary

- Hypothesis tests:

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_A : \theta \neq \theta_0$$

- Use p-values to determine whether to
  - reject the null hypothesis or
  - fail to reject the null hypothesis.
- More assessment is required to determine if other model assumptions hold.

Correspondence between p-values and confidence intervals

STAT 587 (Engineering)
Iowa State University

October 2, 2020

p-values and confidence intervals

From the ASA statement on p-values:

> *a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.*

A $100(1-a)\%$ confidence interval contains the true value of the parameter in $100(1-a)\%$ of the intervals constructed using the procedure.

Both are based on the sampling distribution.

Let $H_0 : \theta = \theta_0$,

- if p-value $< a$, then $100(1-a)\%$ CI will not contain θ_0 but
- if p-value $> a$, then $100(1-a)\%$ CI will contain θ_0.

Normal model

Let $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ with $H_0 : \mu = \mu_0 = 1.5$.

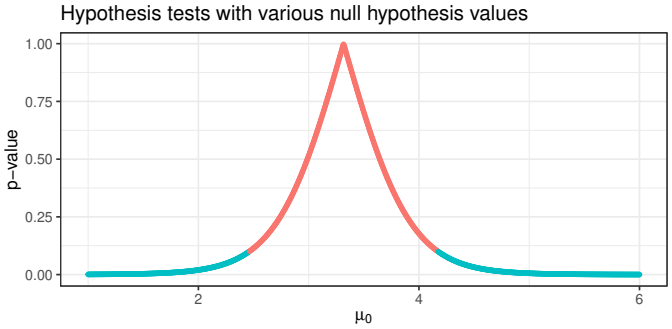
```
y = rnorm(10, mean = 3, sd = 1.5)
a = 0.05
t = t.test(y, mu = mu0, conf.level = 1-a)
t$p.value

[1] 0.003684087

round(as.numeric(t$conf.int),2)

[1] 2.26 4.37
```

```
a = 0.001
t = t.test(y, mu = mu0, conf.level = 1-a)
t$p.value

[1] 0.003684087

round(as.numeric(t$conf.int),2)

[1] 1.08 5.55
```

Explanation

Values for μ_0 that fail to reject H_0 at significance level a are precisely the $100(1-a)\%$ confidence interval.

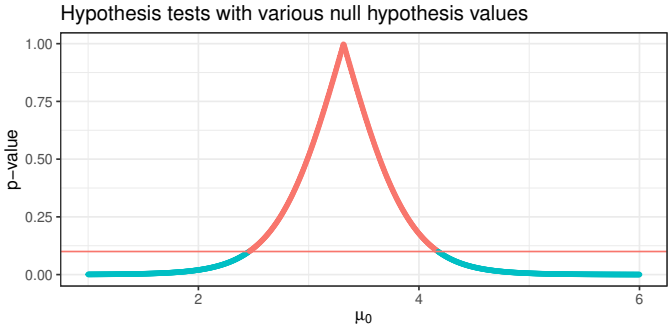
```
a = 0.1
ci = t.test(y, conf.level = 1-a)$conf.int; round(as.numeric(ci),2)

[1] 2.46 4.17
```



Explanation

Values for μ_0 that fail to reject H_0 at significance level a are precisely the $100(1-a)\%$ confidence interval.

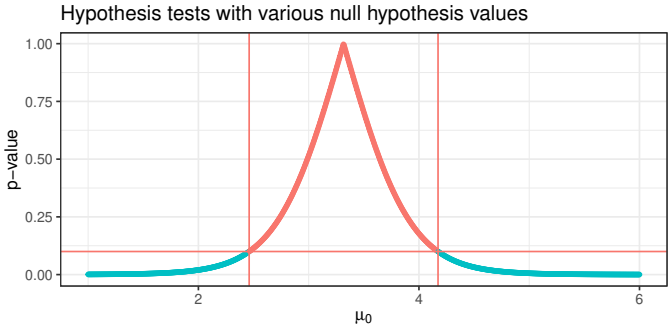
```
a = 0.1
ci = t.test(y, conf.level = 1-a)$conf.int; round(as.numeric(ci),2)

[1] 2.46 4.17
```



Explanation

Values for μ_0 that fail to reject H_0 at significance level a are precisely the $100(1-a)\%$ confidence interval.

```
a = 0.1
ci = t.test(y, conf.level = 1-a)$conf.int; round(as.numeric(ci),2)

[1] 2.46 4.17
```



Importance

The population mean was significantly different than 1.5 ($p = 0.004$).

A 90% confidence interval for the population mean was (2.46, 4.17).

From the second statement, you know

- the p-value is less than 0.1 for any value outside the interval,
- a range of reasonable values for the population mean is given by the interval, and
- a measure of uncertainty given by the interval width and confidence level.

Hypothesis test for a normal mean

for a normal mean

STAT 587 (Engineering)
Iowa State University

September 30, 2020

Statistical hypothesis testing

A hypothesis test consists of two hypotheses:

- null hypothesis (H_0) and
- an alternative hypothesis (H_A)

which make a claim about parameters in a model and a decision to either

- reject the null hypothesis or
- fail to reject the null hypothesis.

Normal model

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then typical hypotheses about the mean are

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_A : \mu \neq \mu_0$$

or

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_A : \mu > \mu_0$$

or

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_A : \mu < \mu_0$$

t-statistic

Then

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}$$

has a t_{n-1} distribution when H_0 is true.
The as or more extreme region is determined by the alternative hypothesis.

$$H_A : \mu < \mu_0 \implies T \leq t$$

or

$$H_A : \mu > \mu_0 \implies T \geq t$$

or

$$H_A : \mu \neq \mu_0 \implies |T| \geq |t|$$

where $T \sim t_{n-1}$.

Example data

Suppose we assume $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$ with $H_0 : \mu = 3$ and we observe

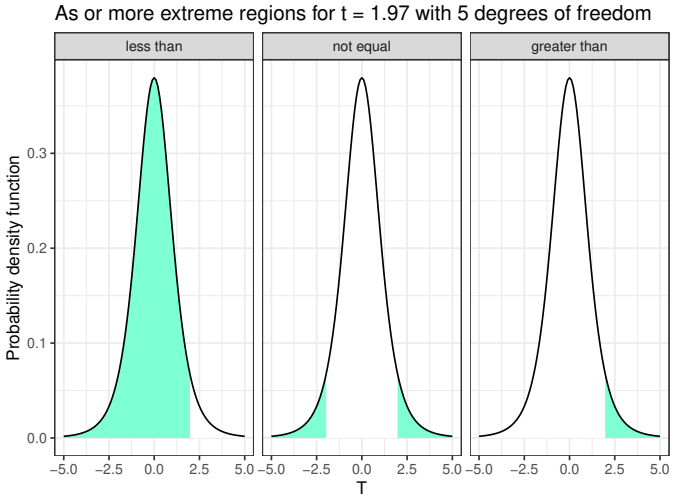
$$n = 6,\ \overline{y} = 6.3,\ \text{and}\ s = 4.1.$$

Then we can calculate

$$t = 1.97$$

which has a t_5 distribution if the null hypothesis is true.

as or more extreme regions



R Calculation

$$H_A : \mu < 3$$

```
t.test(y, mu = mu0, alternative = "less")$p.value
```

```
[1] 0.9461974
```

$$H_A : \mu > 3$$

```
t.test(y, mu = mu0, alternative = "greater")$p.value
```

```
[1] 0.05380256
```

$$H_A : \mu \neq 3$$

```
t.test(y, mu = mu0, alternative = "two.sided")$p.value
```

```
[1] 0.1076051
```

Interpretation

The null hypothesis is a model. For example,

$$H_0 : Y_i \overset{ind}{\sim} N(\mu_0, \sigma^2)$$

if we reject H_0, then we are saying the data are incompatible with this model.

So, possibly

- the Y_i are not independent or
- they don't have a common σ^2 or
- they aren't normally distributed or
- $\mu \neq \mu_0$.

Quality control example

An I-beam manufacturing facility has a design specification for I-beam thickness of 3 millimeters. During manufacturing a random sample of I-beams are taken from the line and their thickness is measured.

```
y
```

```
 [1] 12.04 11.98 11.97 12.12 11.90 12.05 12.14 12.13 12.18 12.23 12.03 12.03
```

```
t.test(y, mu = 12)
```

```
One Sample t-test

data:  y
t = 2.4213, df = 11, p-value = 0.03393
alternative hypothesis: true mean is not equal to 12
95 percent confidence interval:
 12.00607 12.12727
sample estimates:
mean of x
 12.06667
```

The small p-value suggests the data may be incompatible with the model $Y_i \overset{ind}{\sim} N(12, \sigma^2)$.

Summary

- Hypothesis tests for normal mean:

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_A : \mu \neq \mu_0$$

- Use p-values to determine whether to
  - reject the null hypothesis or
  - fail to reject the null hypothesis.
- More assessment is required to determine if other model assumptions hold.

I07 - Posterior model probability

STAT 587 (Engineering)
Iowa State University

October 4, 2020

One-sided alternative hypotheses

For "one-sided alternative hypotheses" just calculate posterior probabilities.

For example, with hypotheses

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_A : \theta > \theta_0$$

Calculate

$$p(H_0|y) = P(\theta \leq \theta_0|y)$$

and

$$p(H_A|y) = P(\theta > \theta_0|y).$$

Posterior probabilities

Let $Y \sim Bin(n, \theta)$ with hypotheses

$$H_0 : \theta \leq 0.5 \qquad \text{and} \qquad H_A : \theta > 0.5.$$

Assume $\theta \sim Unif(0, 1)$ and obtain the posterior i.e.

$$\theta | y \sim Be(1 + y, 1 + n - y).$$

Then calculate

$$p(H_0|y) = P(\theta \leq 0.5|y) = 1 - p(H_A|y).$$

```
n = 10
y = 3
probH0 = pbeta(0.5, 1+y, 1+n-y)
probH0   # p(H_0|y)

[1] 0.8867188

1-probH0 # p(H_A|y)
```

Posterior model probabilities

Calculate the posterior model probabilities over some set of J models i.e,

$$p(M_j|y) = \frac{p(y|M_j)p(M_j)}{p(y)} = \frac{p(y|M_j)p(M_j)}{\sum_{k=1}^{J} p(y|M_k)p(M_k)}.$$

In order to accomplish this, we need to determine

- prior model probabilities:

  $$p(M_j) \quad \text{for all} \quad j = 1, \ldots, J$$

  and

- priors over parameters in each model:

  $$p(y|M_j) = \int p(y|\theta)p(\theta|M_j)d\theta.$$

Prior predictive distribution

The prior predictive distribution for model M_j is

$$p(y|M_j) = \int p(y|\theta)p(\theta|M_j)d\theta.$$

For example, let

$$y|\mu, M_j \sim N(\mu, 1)$$

and

$$\mu|M_j \sim N(0, C),$$

then

$$y|M_j \sim N(0, 1 + C).$$

Bayes Factor

In the context of a null hypothesis (H_0) and an alternative hypothesis (H_A) we have

$$\begin{aligned} p(H_0|y) &= \frac{p(y|H_0)p(H_0)}{p(y|H_0)p(H_0)+p(y|H_A)p(H_A)} \\ &= \left[1 + \frac{p(y|H_A)}{p(y|H_0)}\frac{p(H_A)}{p(H_0)}\right]^{-1} \\ &= \left[1 + BF(H_A : H_0)\frac{p(H_A)}{p(H_0)}\right]^{-1} \end{aligned}$$

where

$$BF(H_A : H_0) = \frac{p(y|H_A)}{p(y|H_0)}$$

is the Bayes Factor for H_A over H_0.

Normal model

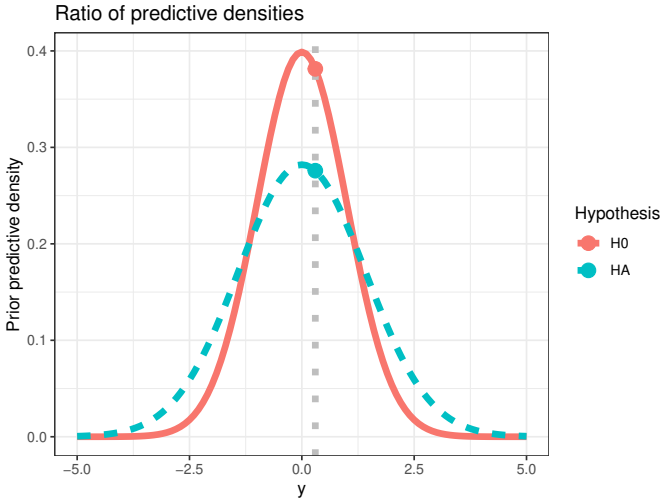
Let $Y \sim N(\mu, 1)$ and $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$.

Assume $p(H_0) = p(H_A)$ and $\mu | H_A \sim N(0, 1)$,
then

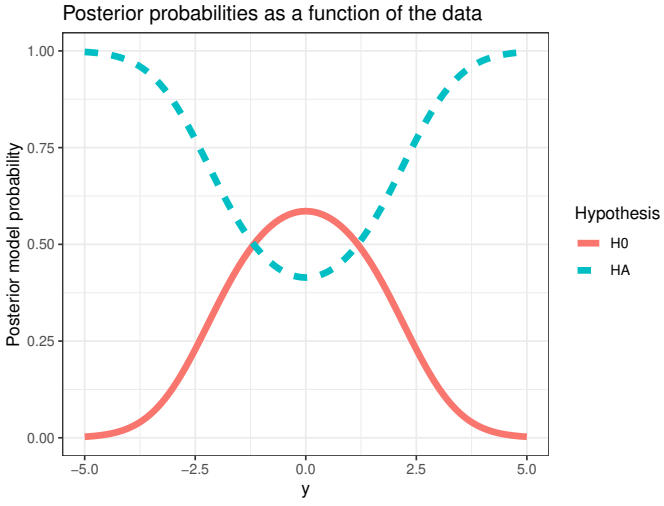
$$\begin{aligned} y|H_0 &\sim N(0, 1) \\ y|H_A &\sim N(0, 2). \end{aligned}$$

```
y = 0.3
probH0 = 1/(1+dnorm(y, 0, sqrt(2))/dnorm(y, 0, 1))
probH0   # p(H_0|y)

[1] 0.5803167

1-probH0 # p(H_A|y)

[1] 0.4196833
```

Ratio of predictive densities



Normal model



Prior impact

Let $Y \sim N(\mu, 1)$ and $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$.

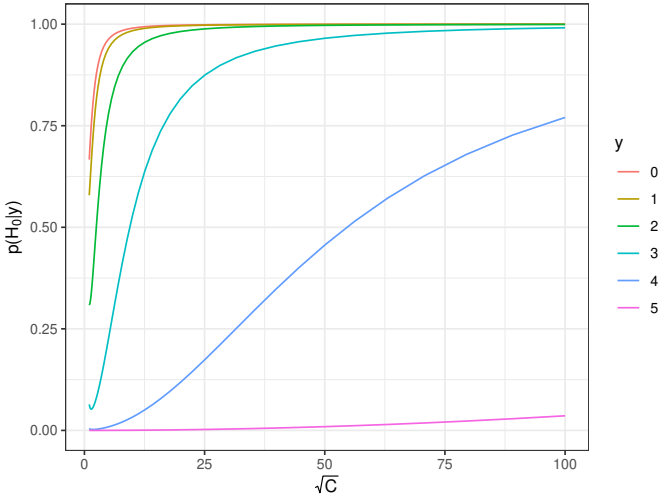
Assume $p(H_0) = p(H_A)$ and $\mu | H_A \sim N(0, C)$, then

$$\begin{aligned} y|H_0 &\sim N(0, 1) \\ y|H_A &\sim N(0, 1 + C) \end{aligned}$$

and

$$p(H_0|y) = \left[1 + \frac{p(y|H_A)}{p(y|H_0)}\right]^{-1}.$$

Prior impact



Interpretation

Since posterior model probabilities depend on the prior predictive distribution

$$p(y|M_j) = \int p(y|\theta)p(\theta|M_j)d\theta$$

posterior model probabilities tell you which model does a better job of prediction and priors, $p(\theta|M_j)$, must be informative.

Do pvalues and posterior probabilities agree?

Suppose $Y \sim Bin(n, \theta)$ and we have the hypotheses $H_0 : \theta = 0.5$ and $H_A : \theta \neq 0.5$ We observe $n = 10,000$ and $y = 4,900$ and find the p-valueis

$$p\text{-value} \approx 2P(Y \leq 4900) = 0.0466$$

so we would reject H_0 at the 0.05 level.

If we assume $p(H_0) = p(H_A) = 0.5$ and $\theta|H_A \sim Unif(0, 1)$, then the posterior probability of H_0, is

$$p(H_0|y) \approx \frac{1}{1 + 1/10.8} = 0.96,$$

so the probability of H_0 being true is 96%.

It appears the posterior probability of H_0 and p-value completely disagree!

Jeffrey-Lindley Paradox

The Jeffrey-Lindley Paradox concerns a situation when comparing two hypotheses H_0 and H_1 given data y and find

- a frequentist test result is significant leading to rejection of H_0, but
- the posterior probability of H_0 is high.

This can happen when

- the effect size is small,
- n is large,
- H_0 is relatively precise,
- H_1 is relative diffuse, and
- the prior model odds is ≈ 1.

No real paradox

p-values:

- a p-value measure how incompatible your data are with the null hypothesis, but
- it says nothing about how incompatible your data are with the alternative hypothesis.

Posterior model probabilities are

- a measure of the (prior) predictive ability of a model relative to the other models, but
- this requires you to have at least two (or more) well-thought out models with informative priors.

Thus, these two statistics provide completely different measures of model adequecy.

Summary

- Use posterior probabilities for one-sided alternative hypotheses.
- Posterior model probabilities evaluate relative predictive ability.

I08 - Comparing probabilities

STAT 587 (Engineering)
Iowa State University

October 4, 2020

One probability

Consider the model $Y \sim Bin(n, \theta)$.

We have discussed a number of statistical procedures to draw inferences about θ:

- Frequentist: based on (asymptotic) distribution of Y/n
  - p-value for test of $H_0 : \theta = \theta_0$,
  - confidence interval for θ,
- Bayesian: based on posterior for θ
  - credible interval for θ,
  - posterior model probability, e.g. $p(H_0|y)$, and
  - posterior probability statements, e.g. $P(\theta < \theta_0|y)$.

Now, we will consider what happens when we have multiple θs.

Two probabilities

Consider the model

$$Y_g \overset{ind}{\sim} Bin(n_g, \theta_g)$$

for $g = 1, 2$ and you are interested in the relationship between θ_1 and θ_2.

- Frequentist: based on asymptotic distribution of $\frac{Y_1}{n_1} - \frac{Y_2}{n_2}$:
  - p-value for a hypothesis test, e.g. $H_0 : \theta_1 = \theta_2$,
  - confidence interval for $\theta_1 - \theta_2$,
- Bayesian: based on posterior distribution of $\theta_1 - \theta_2$:
  - credible interval for θ_1, θ_2,
  - posterior model probability, e.g. $p(H_0|y)$, and
  - probability statements, e.g. $P(\theta_1 < \theta_2|y)$.

where $y = (y_1, y_2)$.

Data example

Suppose you have two manufacturing processes and you are interested in which process has the larger probability of being within the specifications.

So you run the two processes and record the number of successful products produced:

- Process 1: 135 successful products out of 140 attempts
- Process 2: 216 successful products out of 230 attempts

In R, you can code this as two vectors:

```
successes = c(135,216)
attempts  = c(140,230)
```

or, better yet, as a data.frame:

```
d = data.frame(process   = factor(1:2),
               successes = successes,
               attempts  = attempts)
```

p-values and confidence intervals

Because there is no indication that you expect one of the two manufacturing processes to have a higher probability, you should perform a two-sided hypothesis test, i.e.

- $H_0 : \theta_1 = \theta_2$
- $H_A : \theta_1 \neq \theta_2$

and calculate a two-sided confidence interval for $\theta_1 - \theta_2$.

```
prop.test(d$successes, d$attempts)


2-sample test for equality of proportions with continuity correction

data:  d$successes out of d$attempts
X-squared = 0.67305, df = 1, p-value = 0.412
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.02417591  0.07448647
sample estimates:
   prop 1    prop 2
0.9642857 0.9391304
```

Bayesian analysis

Assume

$$Y_g \overset{ind}{\sim} Bin(n_g, \theta_g)$$

and

$$\theta_g \overset{ind}{\sim} Be(1, 1).$$

Then the posterior is

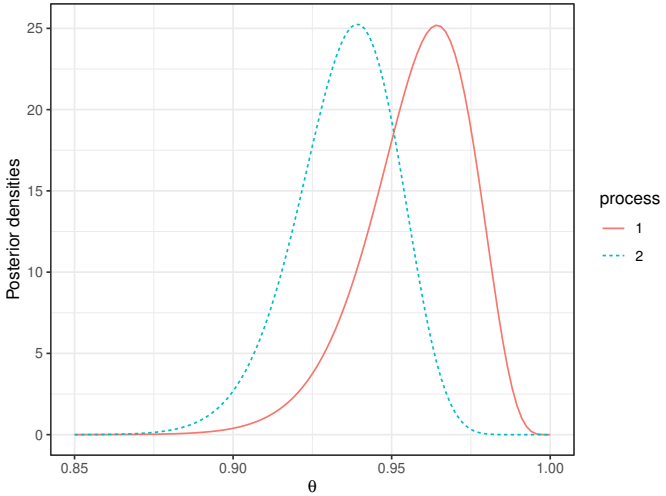
$$\theta_g | y \overset{ind}{\sim} Be(1 + y_g, 1 + n_g - y_g).$$

From this we can compute

$$P(\theta_1 < \theta_2 | y) = P(\theta_1 - \theta_2 < 0 | y)$$

and a credible interval for $\theta_1 - \theta_2$ by simulating values from the posterior and computing $\theta_1 - \theta_2$.

Posteriors


Posterior densities
25
20
15
10
5
0
0.85
0.90
0.95
1.00
θ
process
1
2

Credible interval for the difference

To obtain statistical inference on the difference, we draw samples from the posterior and then calculate the difference:

```
n      <- 1e5
theta1 <- rbeta(n, 1+d$success[1], 1+d$attempts[1] - d$success[1])
theta2 <- rbeta(n, 1+d$success[2], 1+d$attempts[2] - d$success[2])
diff   <- theta1 - theta2

# Bayes estimate for the difference
mean(diff)

[1] 0.02235018

# Estimated 95% equal-tail credible interval
quantile(diff, c(.025,.975))

       2.5%       97.5%
-0.02489203  0.06739588

# Estimate of the probability that theta1 is less than theta2
mean(diff < 0)

[1] 0.16391
```

Multiple probabilities

Now, let's consider the more general problem of

$$Y_g \overset{ind}{\sim} Bin(n_g, \theta_g)$$

for $g = 1, 2, \ldots, G$ and you are interested in the relationship amongst the θ_g.

We can perform the following statistical procedures:

- Frequentist: based on distribution of $Y_1, \ldots, Y_G$
  - p-value for test of $H_0 : \theta_g = \theta$ for all g,
  - p-value for test of $H_0 : \theta_g = \theta_{g'}$,
  - confidence interval for $\theta_g - \theta_{g'}$,
- Bayesian: based on posterior for $\theta_1, \ldots, \theta_G$:
  - credible interval for $\theta_g - \theta_{g'}$,
  - posterior model probability, e.g. $p(H_0|y)$, and
  - probability statements, e.g. $P(\theta_g < \theta_{g'}|y)$.

where g and $'g$ represent different values.

Data example

Suppose you have three manufacturing processes and you are interested in which process has the larger probability of being within the specifications.

So you run the three processes and record the number of successful products produced:

- Process 1: 135 successful products out of 140 attempts
- Process 2: 216 successful products out of 230 attempts
- Process 3: 10 successful products out of 10 attempts

In R, you can code this as two vectors:

```
successes = c(135,216,10)
attempts  = c(140,230,10)
```

or, better yet, as a data.frame:

```
d = data.frame(process   = factor(1:3),
               successes = successes,
               attempts  = attempts)
```

p-values

The default hypothesis test is

$$H_0 : \theta_g = \theta \quad \text{for all} \quad g \quad \text{versus} \quad H_A : \theta_g \neq \theta_{g'} \text{ for some } g, g'$$

```
prop.test(d$successes, d$attempts)

Warning in prop.test(d$successes, d$attempts): Chi-squared approximation may be
incorrect


3-sample test for equality of proportions without continuity correction

data:  d$successes out of d$attempts
X-squared = 1.6999, df = 2, p-value = 0.4274
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.9642857 0.9391304 1.0000000
```

Confidence intervals

Confidence interval for $\theta_1 - \theta_3$:

```
# Need to specify a comparison to get confidence intervals of the difference
prop.test(d$successes[c(1,3)], d$attempts[c(1,3)])$conf.int

Warning in prop.test(d$successes[c(1, 3)], d$attempts[c(1, 3)]):  Chi-squared
approximation may be incorrect

[1] -0.10216886  0.03074029
attr(,"conf.level")
[1] 0.95
```

An alternative test

An alternative test for equality amongst the proportions uses `chisq.test()`.

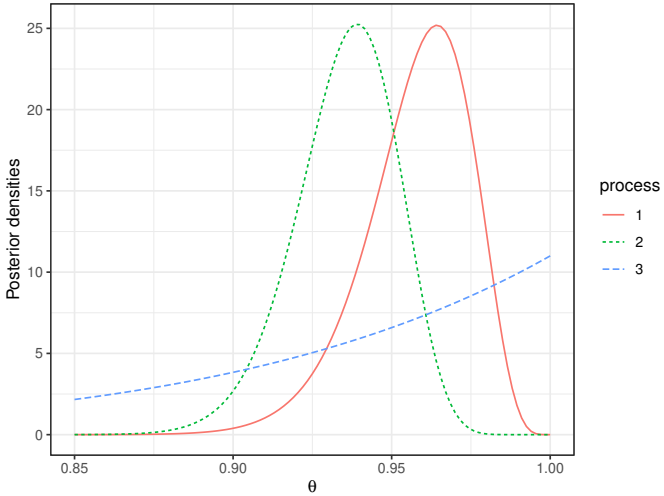
```
d$failures <- d$attempts - d$successes
chisq.test(d[c("successes","failures")])
```

```
Warning in chisq.test(d[c("successes", "failures")]): Chi-squared approximation
may be incorrect

Pearson's Chi-squared test

data:  d[c("successes", "failures")]
X-squared = 1.6999, df = 2, p-value = 0.4274
```

```
chisq.test(d[c("successes","failures")], simulate.p.value = TRUE)
```

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  d[c("successes", "failures")]
X-squared = 1.6999, df = NA, p-value = 0.4158
```

Posteriors


Posterior densities
25
20
15
10
5
0
0.85
0.90
0.95
1.00
θ
process
1
2
3

Credible interval for differences

To compare the probabilities, we draw samples from the posterior and compare them.

```
posterior_samples <- function(d) {
  data.frame(
    rep = 1:1e5,
    name = paste0("theta", d$process),
    theta = rbeta(1e5, 1+d$successes, 1+d$attempts-d$successes),
    stringsAsFactors = FALSE)
}

draws <- d %>% group_by(process) %>% do(posterior_samples(.)) %>% ungroup() %>%
  select(-process) %>% tidyr::spread(name, theta)

# Estimate of the comparison probabilities
draws %>%
  summarize(`P(theta1>theta2|y)` = mean(draws$theta1 > draws$theta2),
            `P(theta1>theta3|y)` = mean(draws$theta1 > draws$theta3),
            `P(theta2>theta3|y)` = mean(draws$theta2 > draws$theta3)) %>%
  gather(comparison, probability)

# A tibble: 3 x 2
  comparison          probability
  <chr>                     <dbl>
1 P(theta1>theta2|y)        0.840
2 P(theta1>theta3|y)        0.632
3 P(theta2>theta3|y)        0.486
```

Summary

Multiple (independent) binomial proportions

- p-values
- confidence intervals
- posterior densities
- credible intervals
- posterior probabilities

I09 - Comparing means

STAT 587 (Engineering)
Iowa State University

October 9, 2020

One mean

Consider the model $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$. We have discussed a number of statistical procedures to draw inferences about μ:

- Frequentist: based on distribution of $\frac{\overline{Y}-\mu}{s/\sqrt{n}}$
  - p-value for a hypothesis test, e.g. $H_0 : \mu = \mu_0$,
  - confidence interval for μ,
- Bayesian: based on posterior for μ
  - credible interval for μ,
  - posterior model probability, e.g. $p(H_0|y)$, and
  - posterior probabilities, e.g. $P(\mu < \mu_0|y)$.

Now, we will consider what happens when you have multiple μs.

Two means

Consider the model

$$Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma_g^2)$$

for $g = 1, 2$ and $i = 1, \ldots, n_g$. and you are interested in the relationship between μ_1 and μ_2.

- Frequentist: based on distribution of

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

  - p-value for a hypothesis test, e.g. $H_0 : \mu_1 = \mu_2$,
  - confidence interval for $\mu_1 - \mu_2$,

- Bayesian: posterior for μ_1, μ_2, i.e. $p(\mu_1, \mu_2|y)$

  - credible interval for $\mu_1 - \mu_2$,
  - posterior model probability, e.g. $p(H_0|y)$, and
  - probability statements, e.g. $P(\mu_1 < \mu_2|y)$.

where $y = (y_{1,1}, \ldots, y_{1,n_1}, y_{2,1}, \ldots, y_{2,n_2})$.

Data example

Suppose you have two manufacturing processes to produce sensors and you are interested in the average sensitivity of the sensors.

So you run the two processes and record the sensitivity of each sensor in units of mV/V/mm Hg (`http://www.ni.com/white-paper/14860/en/`).
And you have the following summary statistics:

```
# A tibble: 2 x 4
  process     n  mean    sd
  <chr>   <int> <dbl> <dbl>
1 P1         22  7.74  1.87
2 P2         34  9.24  2.26
```

p-values and confidence intervals

Because there is no indication that you have any expectation regarding the sensitivities of process 1 compared to process 2, we will conduct a two-sided two-sample t-test assuming the variances are not equal, i.e.

$$Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma_g^2)$$

and

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_A : \mu_1 \neq \mu_2$$

```
t.test(sensitivity ~ process, data = d2)


Welch Two Sample t-test

data:  sensitivity by process
t = -2.6932, df = 50.649, p-value = 0.009571
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.610398 -0.380530
sample estimates:
mean in group P1 mean in group P2
        7.743761         9.239224
```

Posterior for μ_1, μ_2

Assume

$$Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma_g^2) \quad \text{and} \quad p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto \frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2}.$$

Then

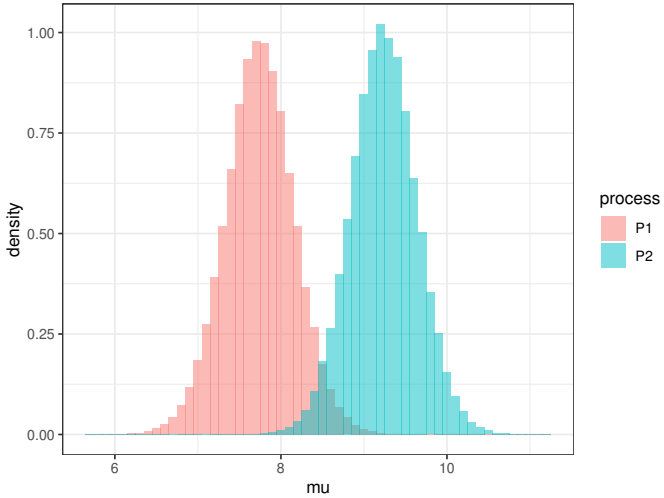
$$\mu_g | y \overset{ind}{\sim} t_{n_g-1}(\overline{y}_g, s_g^2/n_g)$$

and a draw for μ_g can be obtained by taking

$$\overline{y}_g + T_{n_g-1} s_g/\sqrt{n_g}, \quad T_{n_g-1} \overset{ind}{\sim} t_{n_g-1}(0,1).$$

Simulations:

We can use these draws to compare the posteriors



Credible interval for the difference

To obtain statistical inference on the difference, we use the samples and take the difference

```
d3 <- sims %>%
  spread(process, mu) %>%
  mutate(diff = P1-P2)

# Bayes estimate for the difference
mean(d3$diff)

[1] -1.493267

# Estimated 95% equal-tail credible interval
quantile(d3$diff, c(.025,.975))

      2.5%      97.5%
-2.6339752 -0.3483025

# Estimate of the probability that mu1 is larger than mu2
mean(d3$diff > 0)

[1] 0.00591
```

Three or more means

Now, let's consider the more general problem of

$$Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma_g^2)$$

for $g = 1, 2, \ldots, G$ and $i = 1, \ldots, n_g$ and you are interested in the relationship amongst the μ_g.

We can perform the following statistical procedures:

- Frequentist:
  - p-value for test of $H_0 : \mu_g = \mu$ for all g,
  - confidence interval for $\mu_g - \mu_{g'}$,
- Bayesian: based on posterior for $\mu_1, \ldots, \mu_G$
  - credible interval for $\mu_g - \mu_{g'}$,
  - posterior model probability, e.g. $p(H_0|y)$, and
  - probability statements, e.g. $P(\mu_g < \mu_{g'}|y)$

where g and g' are two different groups.

Data example

Suppose you have three manufacturing processes to produce sensors and you are interested in the average sensitivity of the sensors.

So you run the three processes and record the sensitivity of each sensor in units of mV/V/mm Hg (`http://www.ni.com/white-paper/14860/en/`). And you have the following summary statistics:

```
# A tibble: 3 x 4
  process     n  mean    sd
  <chr>   <int> <dbl> <dbl>
1 P1         22  7.74  1.87
2 P2         34  9.24  2.26
3 P3          7 10.8   1.96
```

p-values

When there are lots of means, the first null hypothesis is typically

$$H_0 : \mu_g = \mu \, \forall \, g$$

```
oneway.test(sensitivity ~ process, data = d)

One-way analysis of means (not assuming equal variances)

data:  sensitivity and process
F = 7.6287, num df = 2.000, denom df = 17.418, p-value = 0.004174
```

Pairwise differences

Then we typically look at pairwise differences:

```
pairwise.t.test(d$sensitivity,
                d$process,
                pool.sd = FALSE,
                p.adjust.method = "none")

Pairwise comparisons using t tests with non-pooled SD

data:  d$sensitivity and d$process

   P1     P2
P2 0.0096 -
P3 0.0045 0.0870

P value adjustment method: none
```

Posteriors for μ

When

$$Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma_g^2),$$

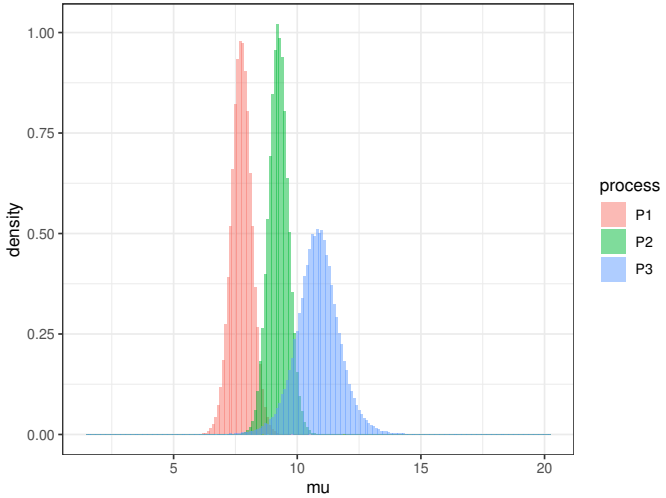
we have

$$\mu_g | y \overset{ind}{\sim} t_{n_g-1}(\overline{y}_g, s_g^2/n_g)$$

and that a draw for μ_g can be obtained by taking

$$\overline{y}_g + T_{n_g-1} s_g/\sqrt{n_g}, \quad T_{n_g-1} \overset{ind}{\sim} t_{n_g-1}(0,1).$$

Compare posteriors



Credible intervals for differences

Use the simulations to calculate posterior probabilities and credible intervals for differences.

```
# Estimate of the probability that one mean is larger than another
sims %>%
  spread(process, mu) %>%
  mutate(`mu1-mu2` = P1-P2,
           `mu1-mu3` = P1-P3,
           `mu2-mu3` = P2-P3) %>%
  select(`mu1-mu2`,`mu1-mu3`,`mu2-mu3`) %>%
  gather(comparison, diff) %>%
  group_by(comparison) %>%
  summarize(probability = mean(diff>0) %>% round(4),
            lower = quantile(diff, .025) %>% round(2),
            upper = quantile(diff, .975) %>% round(2)) %>%
  mutate(credible_interval = paste("(",lower,",",upper,")", sep="")) %>%
  select(comparison, probability, credible_interval)

# A tibble: 3 x 3
  comparison probability credible_interval
  <chr>            <dbl> <chr>
1 mu1-mu2         0.0059 (-2.63,-0.35)
2 mu1-mu3         0.0037 (-5.06,-1.11)
3 mu2-mu3         0.0493 (-3.56,0.37)
```

Common variance model

In the model

$$Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma_g^2)$$

we can calculate a p-value for the following null hypothesis:

$$H_0 : \sigma_g = \sigma \quad \text{for all} \quad g$$

```
bartlett.test(sensitivity ~ process, data = d)

Bartlett test of homogeneity of variances

data:  sensitivity by process
Bartlett's K-squared = 0.90949, df = 2, p-value = 0.6346
```

This may give us reason to proceed as if the variances is the same in all groups, i.e.

$$Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma^2).$$

This assumption is common when the number of observations in the groups is small.

Comparing means when the variances are equal

Assuming $Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma^2)$, we can test

$$H_0 : \mu_g = \mu \, \forall \, g$$

```
oneway.test(sensitivity ~ process, data = d, var.equal = TRUE)

One-way analysis of means

data:  sensitivity and process
F = 6.7543, num df = 2, denom df = 60, p-value = 0.002261
```

Then we typically look at pairwise differences, i.e. $H_0 : \mu_g = \mu_{g'}$.

```
pairwise.t.test(d$sensitivity, d$process, p.adjust.method = "none")

Pairwise comparisons using t tests with pooled SD

data:  d$sensitivity and d$process

   P1     P2
P2 0.0116 -
P3 0.0012 0.0720
```

Posteriors for μ

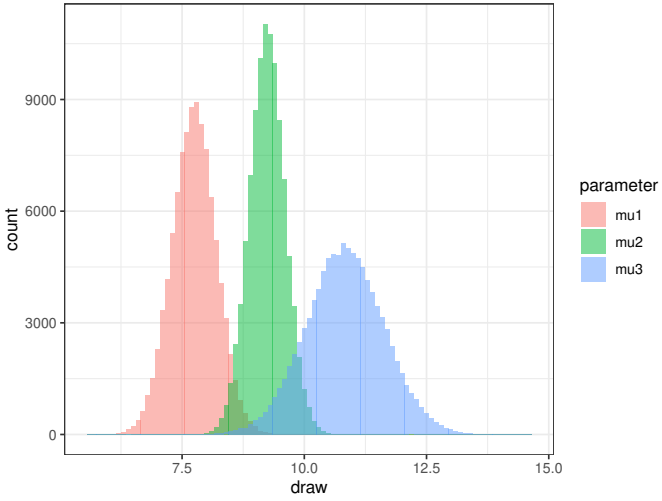
If $Y_{g,i} \overset{ind}{\sim} N(\mu_g, \sigma^2)$ and we use the prior $p(\mu_1, \ldots, \mu_G, \sigma^2) \propto 1/\sigma^2$, then

$$\mu_g|y, \sigma^2 \overset{ind}{\sim} N(\overline{y}_g, \sigma^2/n_g) \quad \sigma^2|y \sim IG\left(\frac{n-G}{2}, \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{g,i} - \overline{y}_g)^2\right)$$

where $n = \sum_{g=1}^{G} n_g$. and thus, we obtain joint samples for μ by performing the following

1. $\sigma^{2(m)} \sim p(\sigma^2|y)$
2. For $g = 1, \ldots, G$, $\mu_g \sim p(\mu_g|y, \sigma^{2(m)})$.

Compare posteriors



Credible interval for the differences

To compare the means, we compare the samples drawn from the posterior.

```
sims %>%
  mutate(`mu1-mu2` = mu1-mu2,
         `mu1-mu3` = mu1-mu3,
         `mu2-mu3` = mu2-mu3) %>%
  select(`mu1-mu2`,`mu1-mu3`,`mu2-mu3`) %>%
  gather(comparison, diff) %>%
  group_by(comparison) %>%
  summarize(probability = mean(diff>0) %>% round(4),
            lower = quantile(diff, .025) %>% round(2),
            upper = quantile(diff, .975) %>% round(2)) %>%
  mutate(credible_interval = paste("(",lower,",",upper,")", sep="")) %>%
  select(comparison, probability, credible_interval)

# A tibble: 3 x 3
  comparison probability credible_interval
  <chr>            <dbl> <chr>
1 mu1-mu2         0.0059 (-2.65,-0.35)
2 mu1-mu3         0.0007 (-4.92,-1.26)
3 mu2-mu3         0.036  (-3.34,0.15)
```

Summary

Multiple (independent) normal means

- p-values
- confidence intervals
- posterior densities
- credible intervals
- posterior probabilities

I10 - Multiple comparisons

STAT 401 (Engineering) - Iowa State University

March 2, 2018

Mice diet effect on lifetimes

Female mice were randomly assigned to six treatment groups to investigate whether restricting dietary intake increases life expectancy. Diet treatments were:

- NP - mice ate unlimited amount of nonpurified, standard diet
- N/N85 - mice fed normally before and after weaning. After weaning, ration was controlled at 85 kcal/wk
- N/R50 - normal diet before weaning and reduced calorie diet (50 kcal/wk) after weaning
- R/R50 - reduced calorie diet of 50 kcal/wk both before and after weaning
- N/R50 lopro - normal diet before weaning, restricted diet (50 kcal/wk) after weaning and dietary protein content decreased with advancing age
- N/R40 - normal diet before weaning and reduced diet (40 Kcal/wk) after weaning.

Exploratory analysis

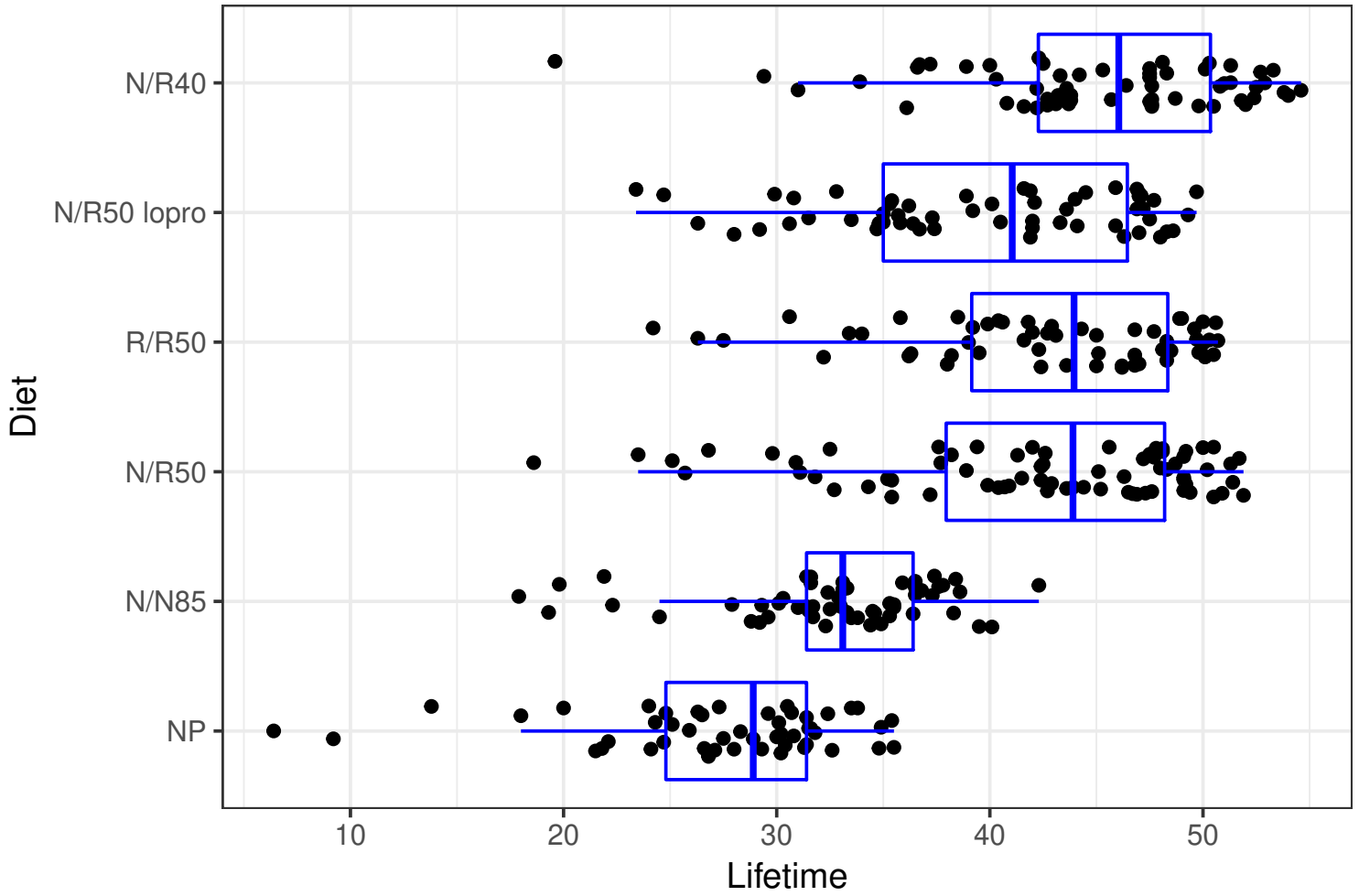
```
library("Sleuth3")
# head(case0501)
summary(case0501)
```

```
    Lifetime       Diet
 Min.   : 6.4   N/N85:57
 1st Qu.:31.8   N/R40:60
 Median :39.5   N/R50:71
 Mean   :38.8   NP   :49
 3rd Qu.:46.9   R/R50:56
 Max.   :54.6   lopro:56
```

```
case0501 <- case0501 %>%
  mutate(Diet = factor(Diet, c("NP","N/N85","N/R50","R/R50","lopro","N/R40")),
         Diet = recode(Diet, lopro = "N/R50 lopro"))
case0501 %>% group_by(Diet) %>% summarize(n=n(), mean = mean(Lifetime), sd = sd(Lifetime))
```

```
# A tibble: 6 x 4
  Diet            n  mean    sd
  <fctr>      <int> <dbl> <dbl>
1 NP             49  27.4  6.13
2 N/N85          57  32.7  5.13
3 N/R50          71  42.3  7.77
4 R/R50          56  42.9  6.68
5 N/R50 lopro    56  39.7  6.99
6 N/R40          60  45.1  6.70
```

```
ggplot(case0501, aes(x=Diet, y=Lifetime)) +
  geom_jitter(width=0.2, height=0) +
  geom_boxplot(fill=NA, color='blue', outlier.color = NA) +
  coord_flip() +
  theme_bw()
```



Are the data compatible with a common mean?

Let Y_{ij} represent the lifetime of mouse j in diet i for $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$. Assume $Y_{ij} \overset{ind}{\sim} N(\mu_i, \sigma^2)$ and calculate a pvalue for $H_0 : \mu_i = \mu$ for all i.

```
bartlett.test(Lifetime ~ Diet, data = case0501)


Bartlett test of homogeneity of variances

data:  Lifetime by Diet
Bartlett's K-squared = 10.996, df = 5, p-value = 0.05146


oneway.test(Lifetime ~ Diet, data = case0501, var.equal = TRUE)


One-way analysis of means

data:  Lifetime and Diet
F = 57.104, num df = 5, denom df = 343, p-value < 2.2e-16


oneway.test(Lifetime ~ Diet, data = case0501, var.equal = FALSE)


One-way analysis of means (not assuming equal variances)

data:  Lifetime and Diet
F = 64.726, num df = 5.00, denom df = 157.84, p-value < 2.2e-16
```

Statistical testing errors

Definition

A type I error occurs when a true null hypothesis is rejected.

Definition

A type II error occurs when a false null hypothesis is not rejected. Power is one minus the type II error probability.

We set our significance level a to control the type I error probability. If we set $a = 0.05$, then we will incorrectly reject a true null hypothesis 5% of the time.

Statistical testing errors

	Truth	
Decision	H_0 true	H_0 false
H_0 not true	Type I error	Correct (power)
H_0 true	Correct	Type II error

Definition

The familywise error rate is the probability of rejecting at least one true null hypothesis.

Type I error for all pairwise comparisons of J groups

How many combinations when choosing 2 items out of J?

$$\binom{J}{2} = \frac{J!}{2!(J-2)!}.$$

If $J = 6$, then there are 15 different comparison of means. If we set $a = 0.05$ as our significance level, then individually each test will only incorrectly reject 5% of the time.

If we have 15 tests and use $a = 0.05$, what is the familywise error rate?

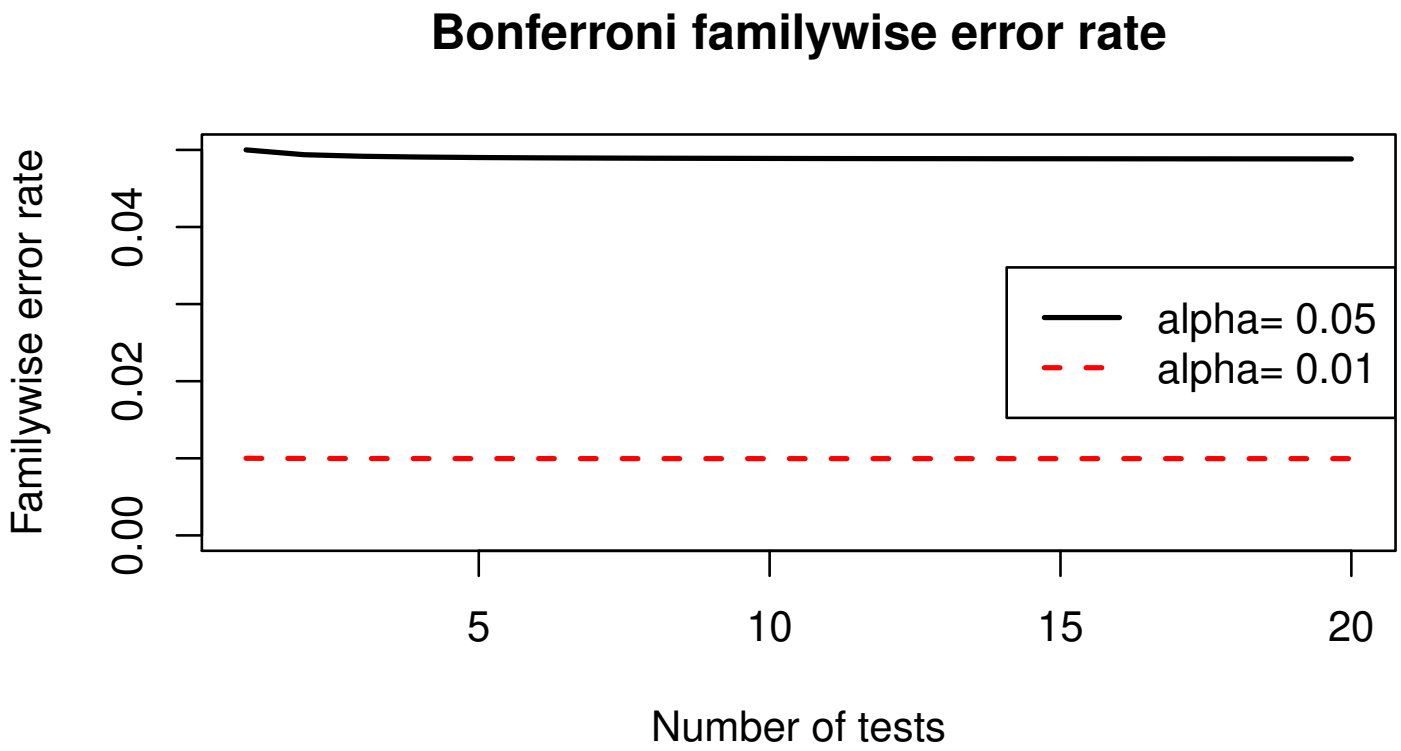
$$1 - (1 - 0.05)^{15} = 1 - (0.95)^{15} = 1 - 0.46 = 0.54$$

So there is a greater than 50% probability of falsely rejecting at least one true null hypothesis!

Bonferroni correction

Definition

If we do m tests and want the familywise error rate to be a, the Bonferroni correction uses a/m for each individual test. The familywise error rate, for independent tests, is $1 - (1 - a/m)^m$.



Pairwise comparisons

If we want to consider all pairwise comparisons of the average lifetimes on the 6 diets, we have 15 tests. In order to maintain a familywise error rate of 0.05, we need a significance level of $0.05/15 = 0.0033333$.

```
pairwise.t.test(case0501$Lifetime, case0501$Diet, p.adjust.method = "none")


Pairwise comparisons using t tests with pooled SD

data:  case0501$Lifetime and case0501$Diet

            NP      N/N85   N/R50 R/R50 N/R50 lopro
N/N85       5.9e-05 -       -     -     -
N/R50       < 2e-16 1.1e-14 -     -     -
R/R50       < 2e-16 8.9e-15 0.622 -     -
N/R50 lopro < 2e-16 5.2e-08 0.029 0.012 -
N/R40       < 2e-16 < 2e-16 0.017 0.073 1.6e-05

P value adjustment method: none
```

Pairwise comparisons

If we want to consider all pairwise comparisons of the average lifetimes on the 6 diets, we have 15 tests. Alternatively, you can let R do the adjusting for you, but now you need to compare with the original significance level a.

```
pairwise.t.test(case0501$Lifetime, case0501$Diet, p.adjust.method = "bonferroni")


Pairwise comparisons using t tests with pooled SD

data:  case0501$Lifetime and case0501$Diet

            NP      N/N85   N/R50   R/R50   N/R50 lopro
N/N85       0.00089 -       -       -       -
N/R50       < 2e-16 1.6e-13 -       -       -
R/R50       < 2e-16 1.3e-13 1.00000 -       -
N/R50 lopro < 2e-16 7.9e-07 0.44018 0.17507 -
N/R40       < 2e-16 < 2e-16 0.24881 1.00000 0.00024

P value adjustment method: bonferroni
```

Comments on the Bonferroni correction

The Bonferroni correction can be used in any situation. In particular, it can be used on unadjusted pvalues reported in an article that has many tests by comparing their pvalues to a/m where m is the number of tests they perform.

The Bonferroni correction is (in general) the **most** conservative multiple comparison adjustment, i.e. it will lead to the least null hypothesis rejections.

Constructing multiple confidence intervals

A $100(1-a)\%$ confidence interval should contain the true value $100(1-a)\%$ of the time when used with different data sets.

An error occurs if the confidence interval does not contain the true value.

Just like the Type I error and familywise error rate, we can ask what is the probability at least one confidence interval does not cover the true value.

The procedures we will talk about for confidence intervals have equivalent approaches for hypothesis testing (pvalues). Within these procedures we still have the equivalence between pvalues and CIs.

Constructing multiple confidence intervals

Confidence interval for the difference between group j and group j':

$$\overline{Y}_j - \overline{Y}_{j'} \pm M\, s_p \sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}$$

where M is a multiplier that depends on the adjustment procedure:

Procedure	M	Use
LSD	$t_{n-J}(1-a/2)$	After significant F-test (no adjustment)
Dunnett	multivariate t	Compare all groups to control
Tukey-Kramer	$q_{J,n-J}(1-a)/\sqrt{2}$	All pairwise comparisons
Scheffé	$\sqrt{(J-1)F_{(J-1,n-J)}(1-a)}$	All contrasts
Bonferroni	$t_{n-J}(1-(a/m)/2)$	m tests (most generic)

Tukey for all pairwise comparisons

```
TukeyHSD(aov(Lifetime ~ Diet, data = case0501))


  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Lifetime ~ Diet, data = case0501)

$Diet
                        diff        lwr        upr     p adj
N/N85-NP           5.2891873  1.5606269  9.0177476 0.0008380
N/R50-NP          14.8951423 11.3405719 18.4497127 0.0000000
R/R50-NP          15.4836735 11.7397556 19.2275913 0.0000000
N/R50 lopro-NP    12.2836735  8.5397556 16.0275913 0.0000000
N/R40-NP          17.7146259 14.0294069 21.3998448 0.0000000
N/R50-N/N85        9.6059550  6.2021702 13.0097399 0.0000000
R/R50-N/N85       10.1944862  6.5934168 13.7955556 0.0000000
N/R50 lopro-N/N85  6.9944862  3.3934168 10.5955556 0.0000008
N/R40-N/N85       12.4254386  8.8854359 15.9654413 0.0000000
R/R50-N/R50        0.5885312 -2.8320696  4.0091319 0.9963976
N/R50 lopro-N/R50 -2.6114688 -6.0320696  0.8091319 0.2460200
N/R40-N/R50        2.8194836 -0.5367684  6.1757356 0.1564608
N/R50 lopro-R/R50 -3.2000000 -6.8169683  0.4169683 0.1167873
N/R40-R/R50        2.2309524 -1.3252222  5.7871269 0.4684413
N/R40-N/R50 lopro  5.4309524  1.8747778  8.9871269 0.0002306
```

False Discovery Rate

Not wanting to make a single mistake is pretty conservative.
In high-throughput fields a more common multiple comparison adjustment is false discovery rate.

Definition

False discovery rate procedures try to control the expected proportion of incorrectly rejected null hypotheses.

How to incorporate multiple comparison adjustments

1. Determine what tests are going to be run (before looking at the data) or what confidence intervals are going to be constructed.
2. Determine which multiple comparison adjustment is the most relevant.
3. Use/state that adjustment and interpret your results.