Stats Foundations

(close to Appendix C Wooldridge)

M Loecher

Estimators

Random Sampling

- If $Y_1, \ldots, Y_2, Y_n$ are independent random variables with a common probability density function $f(y;\theta)$, then $\{Y_1, \ldots, Y_2, Y_n\}$ is said to be a **random sample** from $f(y;\theta)$.
- We also say that the Yi are **independent, identically distributed (or i.i.d.)** random variables from $f(y;\theta)$.

Estimators

- Given a random sample $Y_1, \ldots, Y_2, Y_n$ drawn from a population distribution that depends on an unknown parameter θ, an **estimator** $W = h(Y_1, \ldots, Y_2, Y_n)$ of θ is a rule that assigns each possible outcome of the sample a value of θ.
- A natural estimator of μ is the average of the random sample:

$$\overline{Y} = \frac{1}{n} \cdot \sum_{i=1}^{n} Y_i$$

- $\overline{Y}$ is called the sample average but, while earlier we defined the sample average of a set of numbers as a descriptive statistic, $\overline{Y}$ is now viewed as an estimator !

Expected Values

Generalization of average for any random variable x:

$$\mu = E(x) = \sum_{i=1}^{k} p_i \cdot x_i$$

Example coin flip:

- E(X) = ?
- E(Var(X)) = E($(x - \mu)^2$) = ?

Unbiasedness

- The distribution of an estimator is often called its **sampling distribution**, because this distribution describes the likelihood of various outcomes across different random samples.
- An estimator, W of θ, is an **unbiased estimator** if

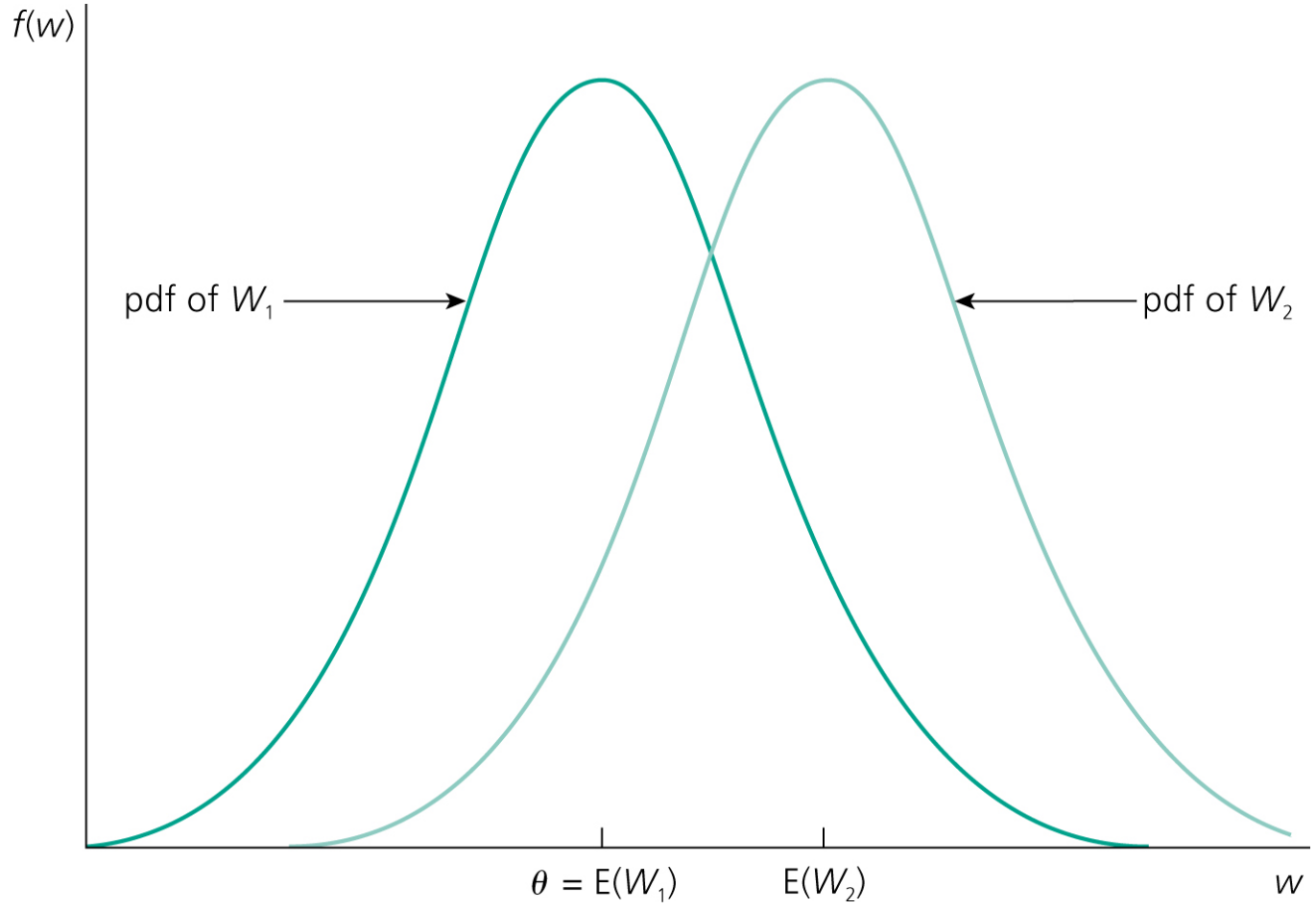
$$E(W) = \theta$$

- If W is a **biased estimator** of θ, its bias is defined as

$$Bias(W) = E(W) - \theta$$

- The sample mean is unbiased: $E(\overline{Y}) = \mu$

(Un)biased Estimators



Sample Variance

- $$\frac{1}{n} \cdot \sum_{i=1}^{n} (Y_i - \mu)^2$$

  is an unbiased estimator for σ^2
- However: $E(\sigma^2) \neq \frac{1}{n} \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2$
- We define the **sample variance** as an unbiased version:

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = E(\sigma^2)$$

Sampling Variance of Estimators

- The variance of an estimator is often called its **sampling variance**. We can show:

$$Var(\overline{Y}) = Var\left(\frac{1}{n} \cdot \sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} Y_i\right)$$
$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(Y_i) = \sigma^2/n$$

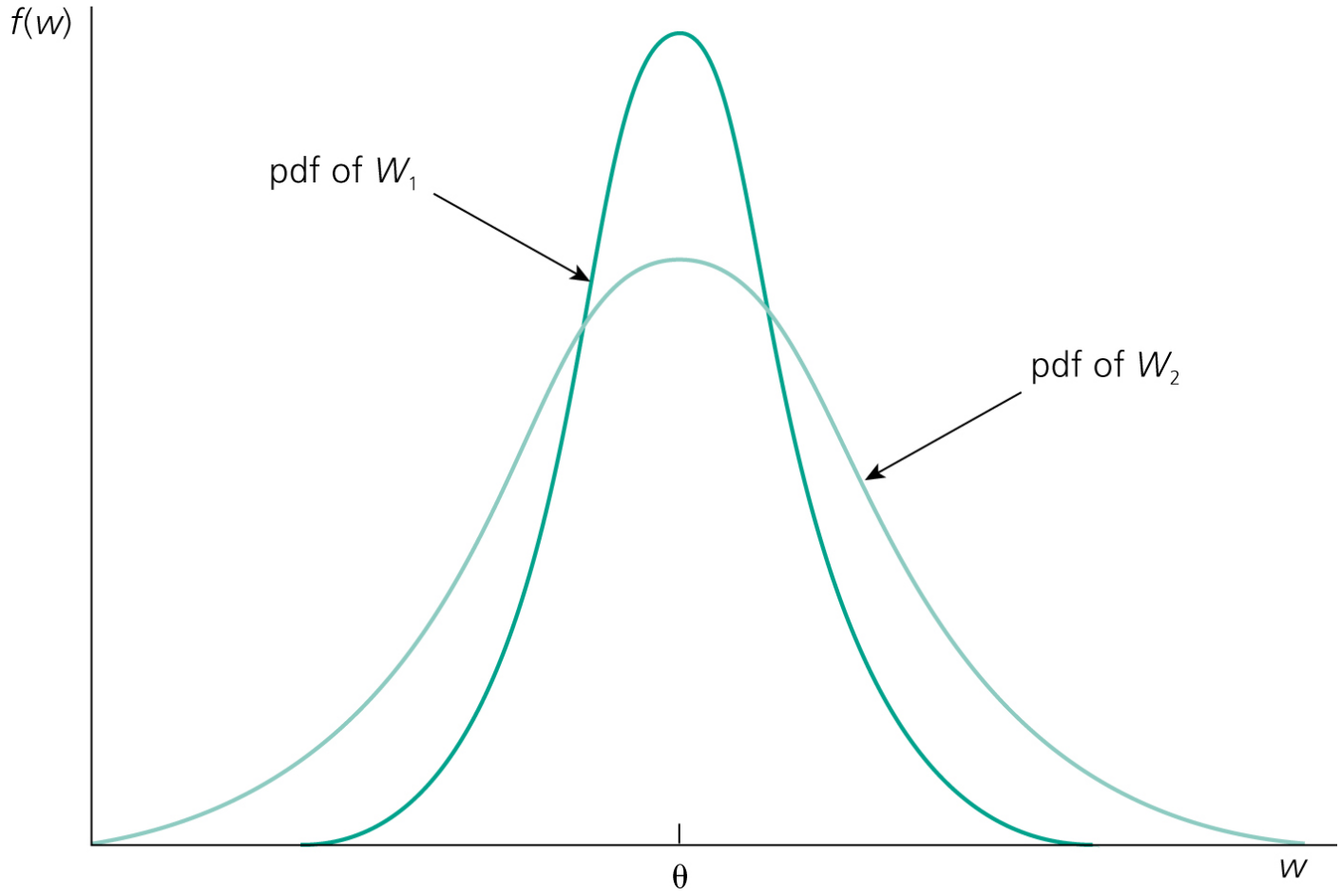
- Hence:

$$\sigma(\bar{Y}) = \sigma/\sqrt{n}$$

which has its own name: the **standard error**

Good Estimators

We prefer the estimator with the smallest variance. This allows us to eliminate certain estimators from consideration.



Efficiency

- If W_1 and W_2 are two unbiased estimators of θ, W_1 is **efficient** relative to W_2 when $\text{Var}(W_1) \leq \text{Var}(W_2)$ for all θ, with strict inequality for at least one value of θ.
- $\overline{Y}$ has the smallest variance among all unbiased estimators that are also linear functions of $Y_1, \ldots, Y_2, Y_n$.
- If we do not restrict our attention to unbiased estimators, then comparing variances is meaningless.
- One way to compare estimators that are not necessarily unbiased is to compute the **mean squared error (MSE)** of the estimators. If W is an estimator of θ, then the MSE of W is defined as $\text{MSE(W)} = E[(W - \theta)^2]$.

Consistency

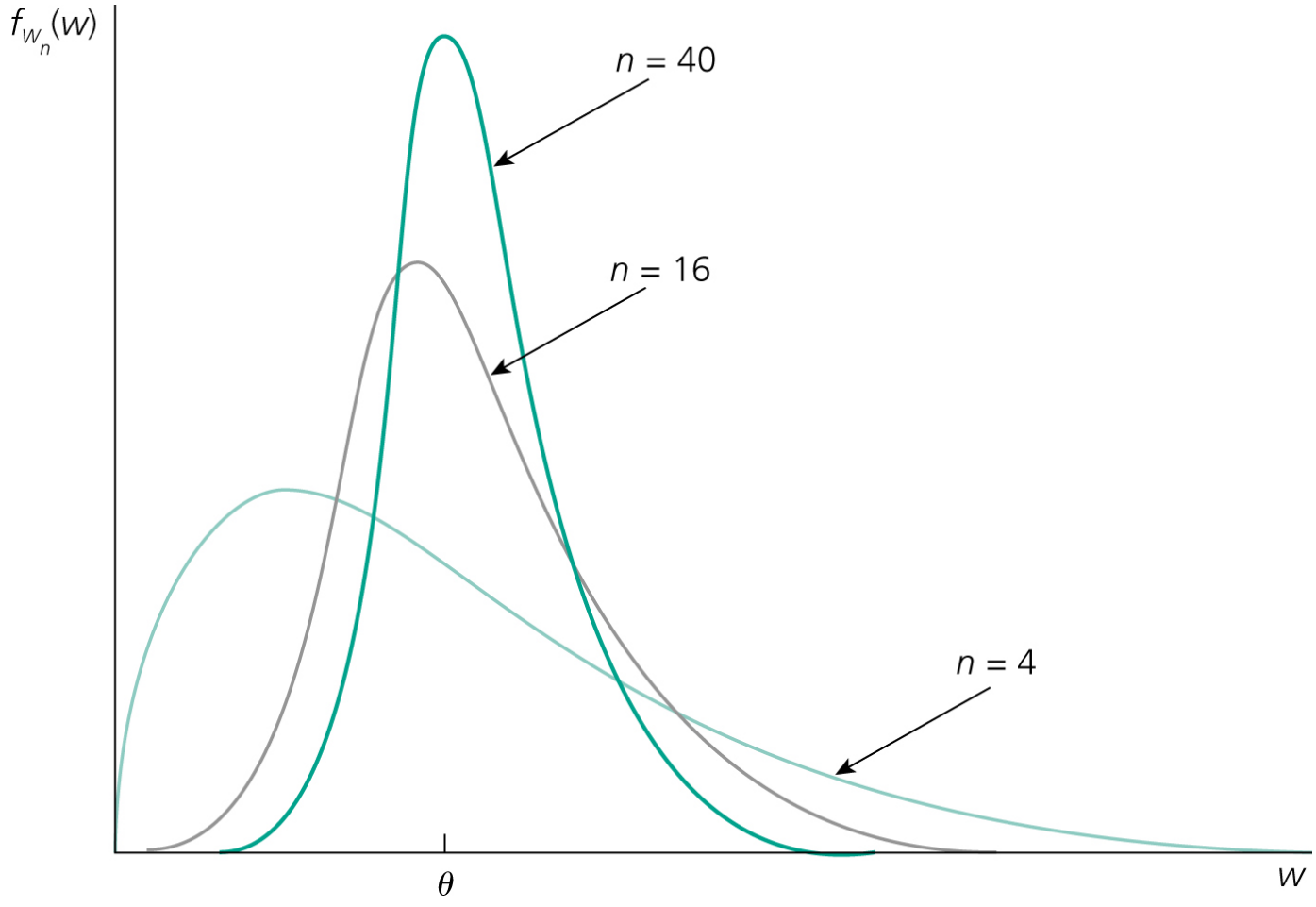
- Let W_n be an estimator of θ based on a sample $Y_1, \ldots, Y_2, Y_n$ of size n. Then, W_n is a **consistent estimator** of θ if for every $\epsilon > 0$

$$P(|W_n - \theta| > \epsilon) \to 0 \text{ as } n \to \infty$$

- When W_n is consistent, we also say that θ is the probability limit of W_n, written as $\text{plim}(\ W_n) = \theta$
- Consistency is a minimal requirement of an estimator used in statistics or econometrics

Consistency

The distribution of W_n becomes more and more concentrated about θ:



Sampling

Review Sampling Distribution

Truth to sample ("Left to right")

Given μ, σ and an **i.i.d. sample of size n** we can make statements about the likelihood of values the sample mean $\bar{x}_n$ will take on, e.g. for $\mu = 96, \sigma = 4, n = 64$:

1. Find the **interval** which contains the sample mean with 99% probability.
2. Find the **cutoff** such that the sample mean lies below with 95% probability.
3. Find the **probability** that the sample mean will be no more than $95.36cm$.

Review, Solution

1. $96 \pm z_{0.995} \cdot 4/\sqrt{64} = 96 \pm 2.575 \cdot 0.5 = [94.7125; 97.2875]$.
2. $96 + z_{0.95} \cdot 4/\sqrt{64} = 96 + 1.65 \cdot 0.5 = 96.825$.
3. $z = \frac{95.36-96}{0.5} = -1.28, \Rightarrow P(z < -1.28) = 0.1$

Inference I (given σ)

Sample to Truth ("Right to Left")

Given $\bar{x}_n, \sigma$ we would like to make statements about likely values of the **true mean**, e.g. for $\bar{x}_{64} = 96.5, \sigma = 4$:

1. Find the **interval** which contains the true mean with 99% probability.
2. Find the **cutoff** such that the true mean lies below with 95% probability.
3. Find the **probability** that the true mean will be no more than $95.86cm$.

Inference I, Solution

$\bar{x}_{64} = 96.5, \sigma = 4$

1. $96.5 \pm z_{0.995} \cdot 4/\sqrt{64} = 96.5 \pm 2.575 \cdot 0.5 = [95.212597.7875]$.
2. $96.5 + z_{0.95} \cdot 4/\sqrt{64} = 96.5 + 1.65 \cdot 0.5 = 97.325$.
3. Another way of saying this is:

   $H_0 : \mu \leq 95.86,\ H_A : \mu > 95.86$

   $z = \frac{96.5-95.86}{0.5} = 1.28, \Rightarrow P(z \geq 1.28) = 0.1$

Inference II (unknown σ)

Sample to Truth ("Right to Left")

Given $\bar{x}_n$ and the sample standard deviation s_n we would like to make statements about likely values of the **true mean**, e.g. for $\bar{x}_{64} = 96.5, s_{64} = 3.8$:

1. Find the **interval** which contains the true mean with 99% probability.
2. Find the **cutoff** such that the true mean lies below with 95% probability.
3. Find the **probability** that the true mean will be no more than $95.88cm$.

Inference II, Solution

$\bar{x}_{64} = 96.5, s_{64} = 3.8$

1. $96.5 \pm t_{0.995,63} \cdot 3.8/\sqrt{64} = 96.5 \pm 2.656 \cdot 0.475 = [95.23888; 97.76112]$.
2. $96.5 + t_{0.95,63} \cdot 3.8/\sqrt{64} = 96.5 + 1.67 \cdot 0.475 = 97.2932$.
3. Another way of saying this is:

   $H_0 : \mu \leq 95.88,\ H_A : \mu > 95.88$

   $z = \frac{96.5-95.88}{0.475} = 1.28, \Rightarrow P(t \geq 1.28) = 0.1$

Inference III (Hypothesis Testing)

The inference above had to with **estimation** which is extremely valuable. But often, we need to make decisions on the plausability of a claim (a "hypothesis") regarding a parameter of interest.

Given $\bar{x}_n, s_n$ we would like to make statements about the plausability of claims regarding the **true mean**, e.g. for $\bar{x}_{64} = 96.5, s_{64} = 3.8$:

1. Is it possible that the true mean is $\mu = 95cm$?
2. Someone claims that the true mean is at least $97.5cm$.
3. Find the **probability** that the true mean will be no more than $95.88cm$.

Law of Large Numbers

Let $Y_1, \ldots, Y_2, Y_n$ be independent, identically distributed random variables with mean μ. Then,

$$\text{plim}(\overline{Y}_n) = \mu$$

$$\Rightarrow \overline{Y}_n \sim N\left(\mu; \sigma/\sqrt{n}\right) = N(\mu; se(\overline{Y}_n))$$

Quiz

$$E(X - a) = ?$$

$$E\left(\frac{X - a}{b}\right) = ?$$

$$Var(X - a) = ?$$

$$Var\left(\frac{X - a}{b}\right) = ?$$

Standardization

We often standardize variables by measuring their deviation from the mean in terms of multiples of standard deviations, which we refer to as the z-score:

$$z = \frac{x - \mu}{\sigma_x}$$

We should think of this ratio in more general terms as

$$z = \frac{\text{Estimate} - \text{Reference}}{\text{Stdev of the Estimate}}$$

That leaves open the possibility to use any uncertainty measure in the denominator such as the standard error!

Central Limit Theorem

Let $\{Y_1, \ldots, Y_2, Y_n\}$ be a random sample with mean μ and variance σ^2. Then

$$Z_n = \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \sim N(0; 1)$$

("has an asymptotic standard normal distribution.")

An important variation is obtained by replacing σ with its consistent estimator S_n:

$$T_n = \frac{\overline{Y}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

Interval containing the sample mean

$$\Rightarrow \overline{Y}_n \sim N\left(\mu; \sigma/\sqrt{n}\right) = N(\mu; se(\overline{Y}_n))$$

$$\Rightarrow P\left(|\overline{Y}_n - \mu| < \mathbf{1.96} \cdot \sigma/\sqrt{n}\right) = \mathbf{0.95}$$

$$\Rightarrow P\left(|\overline{Y}_n - \mu| < \mathbf{z_{1-\alpha/2}} \cdot \sigma/\sqrt{n}\right) = \mathbf{1} - \alpha$$

$$\Rightarrow P\left(\frac{|\overline{Y}_n - \mu|}{\sigma/\sqrt{n}} < \mathbf{z_{1-\alpha/2}}\right) = \mathbf{1} - \alpha$$

For known/given μ, σ the above equation is a statement on how the sample mean fluctuates around the true mean μ.

Confidence Intervals

But if only $\overline{Y}_n$ is known, we can reverse the logic and make statements about the true mean:

With 0.95 $(= 1 - \alpha)$ probability, μ lies no more than 1.96 $(= z_{1-\alpha/2})$ std. errors from the sample mean $\overline{Y}_n$!

And since we do not normally know σ either, we replace it with its estimator $s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2}$ and end up with a different multiplier from the **t distribution**:

$$\Rightarrow P\left(|\overline{Y}_n - \mu| < \mathbf{t_{1-\alpha/2,n-1}} \cdot s/\sqrt{n}\right) = \mathbf{1} - \alpha$$

e.g. for $n = 20, \alpha = 0.05$

$$\Rightarrow P\left(|\overline{Y}_n - \mu| < \mathbf{2.1} \cdot s/\sqrt{n}\right) = \mathbf{0.95}$$

Short hand notation

Let $c_{\alpha/2}$ denote the $100(1-\alpha)$ **percentile** of the t_{n-1} distribution. Define the standard error of $\overline{Y}$ as $\text{se}(\bar{y})$.

Then a $100(1-\alpha)\%$ confidence interval is simply

$$[\bar{y} \pm c_{\alpha/2} \cdot se(\bar{y})]$$

Review

We have encountered two situations even in the best case of independent draws from a normal distribution, $X_i \sim N(\mu, \sigma)$:

- If σ is known, the random variable

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{SE_{\bar{x}}} \sim N(0, 1)$$

- If we estimate σ by the sample stdev $\hat{\sigma}$, the random variable

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} = t_{n-1}$$

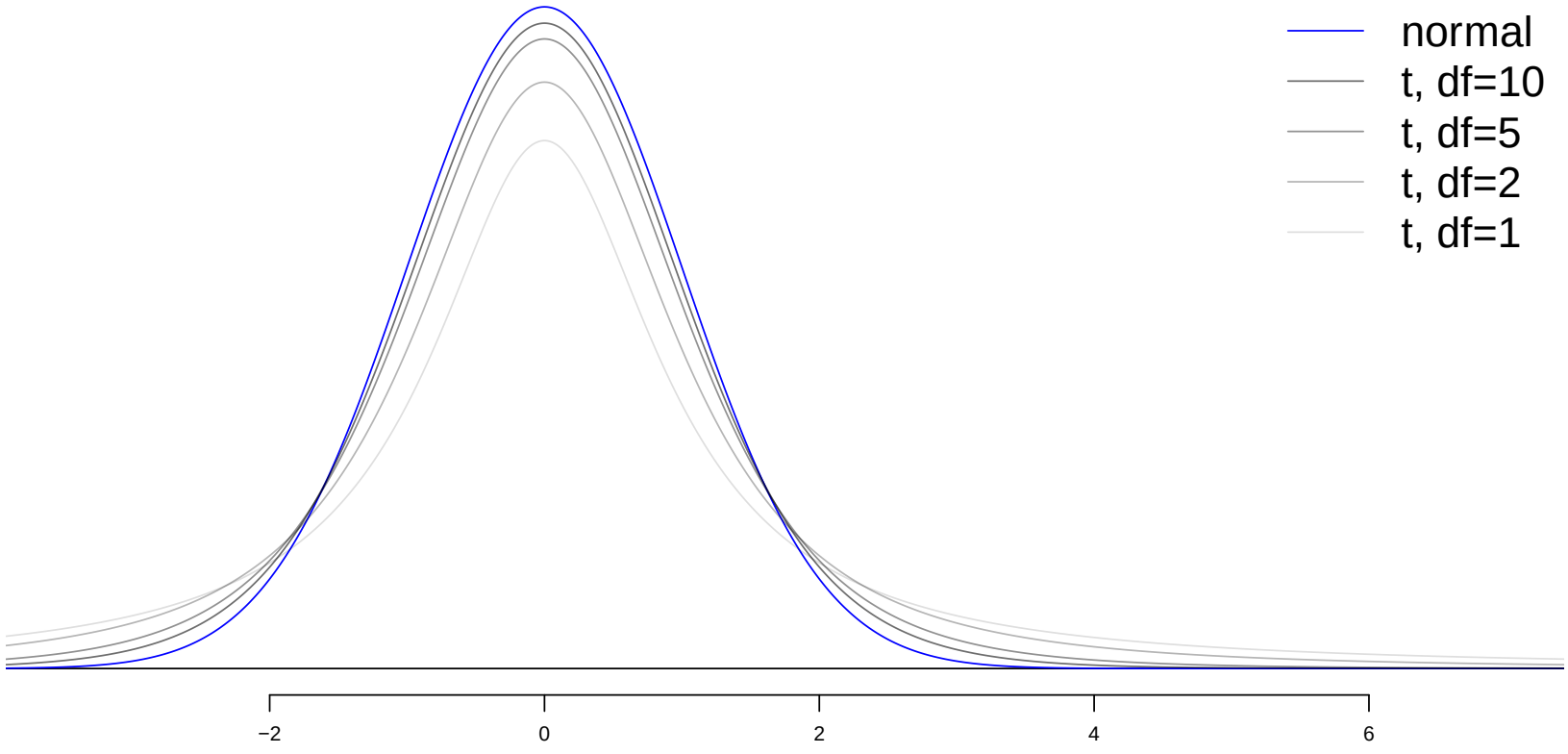
  follows a **t-distribution** with $n - 1$ **degrees of freedom**!

- The sample stdev s is not an unbiased estimator of $\sigma \Rightarrow$ we divide by N-1 instead of N! [1]

[1] $\hat{\sigma}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$

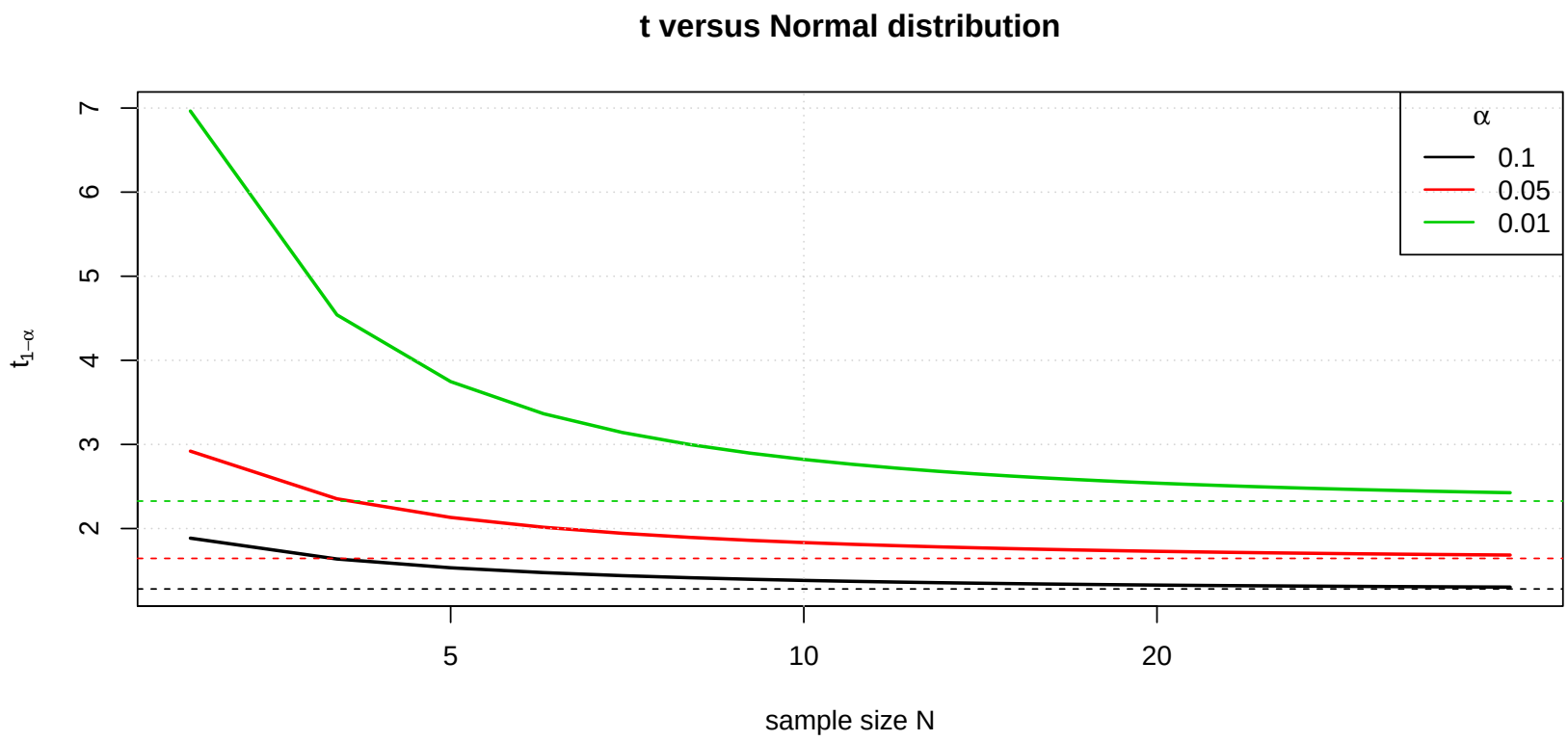
The t-distribution, I

Thicker Tails:



The t-distribution, II

Various combinations of α and N (The horizontal dashed lines are the corresponding values from the normal distribution).



t table

	$1-\alpha$				
DF	**0.9**	**0.95**	**0.975**	**0.99**	**0.995**
3	1.64	2.35	3.18	4.54	5.84
5	1.48	2.02	2.57	3.36	4.03
10	1.37	1.81	2.23	2.76	3.17
15	1.34	1.75	2.13	2.60	2.95
20	1.33	1.72	2.09	2.53	2.85
30	1.31	1.70	2.04	2.46	2.75
z	1.28	1.64	1.96	2.33	2.58

Confidence Intervals:

$$\bar{x} \pm z_{1-\alpha/2} \cdot \sigma/\sqrt{n}, \text{ or}^2 \ \bar{x} \pm t_{(n-1,1-\alpha/2)} \cdot \hat{\sigma}/\sqrt{n}$$

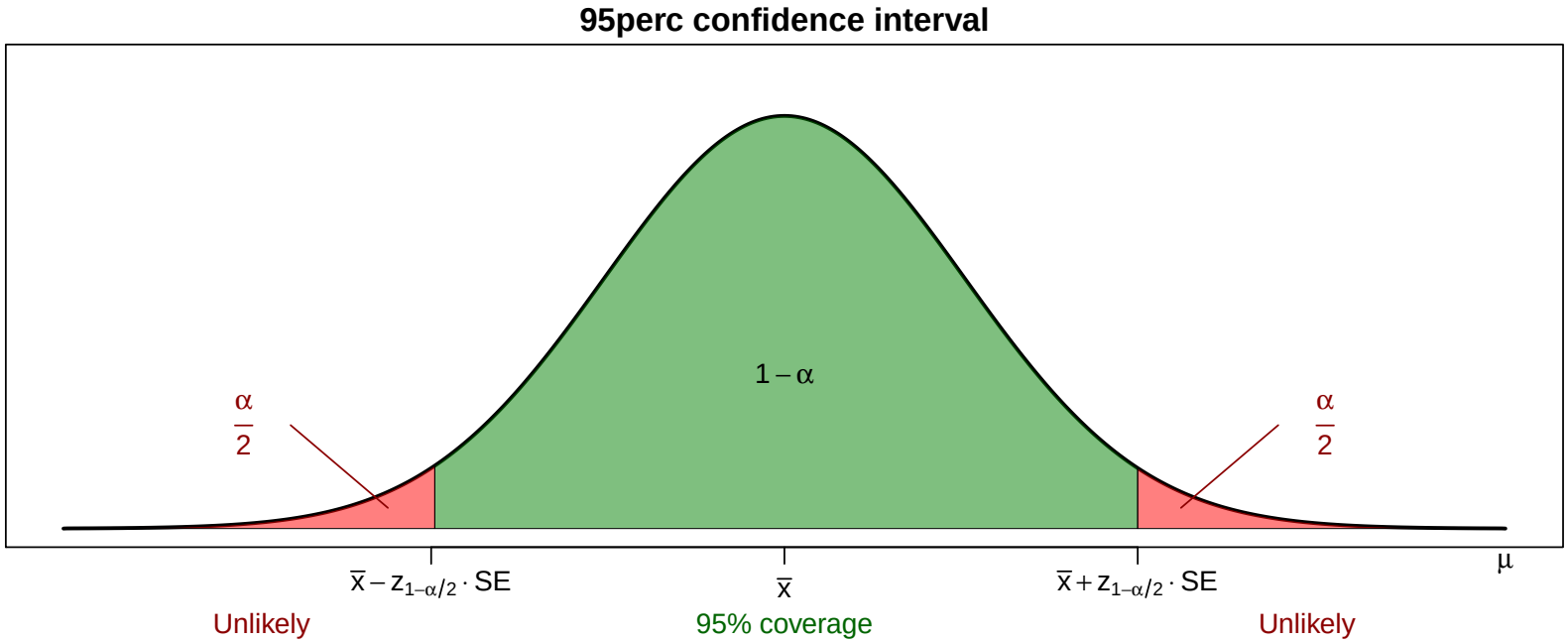
[2] if $\hat{\sigma}$ has to be estimated from the data

Formal Confidence Intervals:

$$P(\mu - z_{1-\alpha/2} \cdot SE < \bar{x} \leq \mu + z_{1-\alpha/2} \cdot SE) = 1 - \alpha$$

$$\Leftrightarrow$$

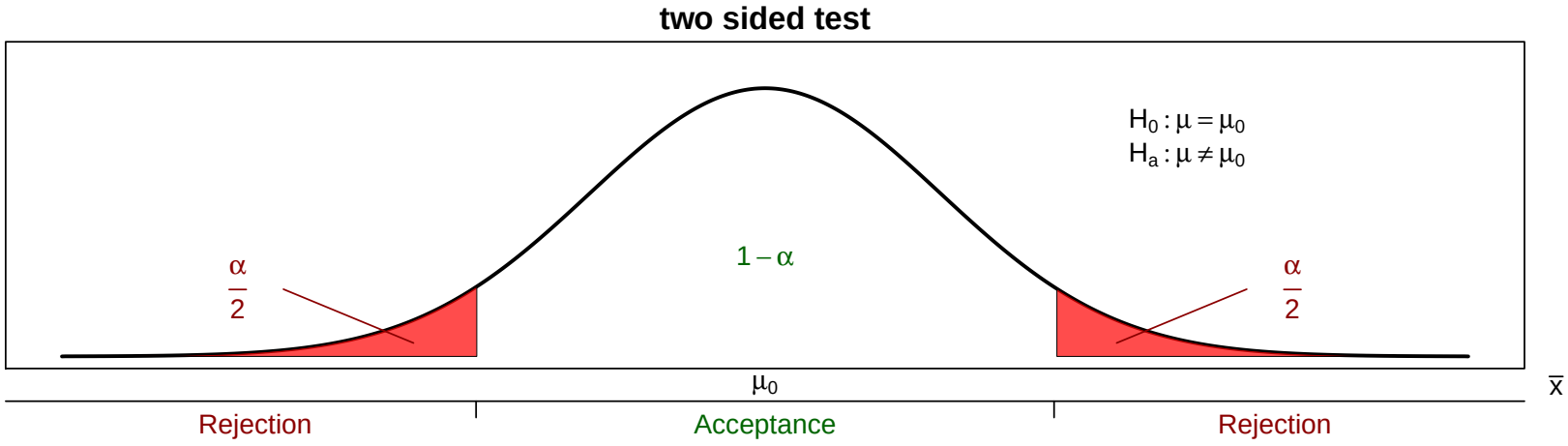
$$P(\bar{x} - z_{1-\alpha/2} \cdot SE \leq \mu < \bar{x} + z_{1-\alpha/2} \cdot SE) = 1 - \alpha$$



Hypothesis Tests

Hypothesis Tests

$$H_0 : \mu = \mu_0, H_a : \mu \neq \mu_0$$



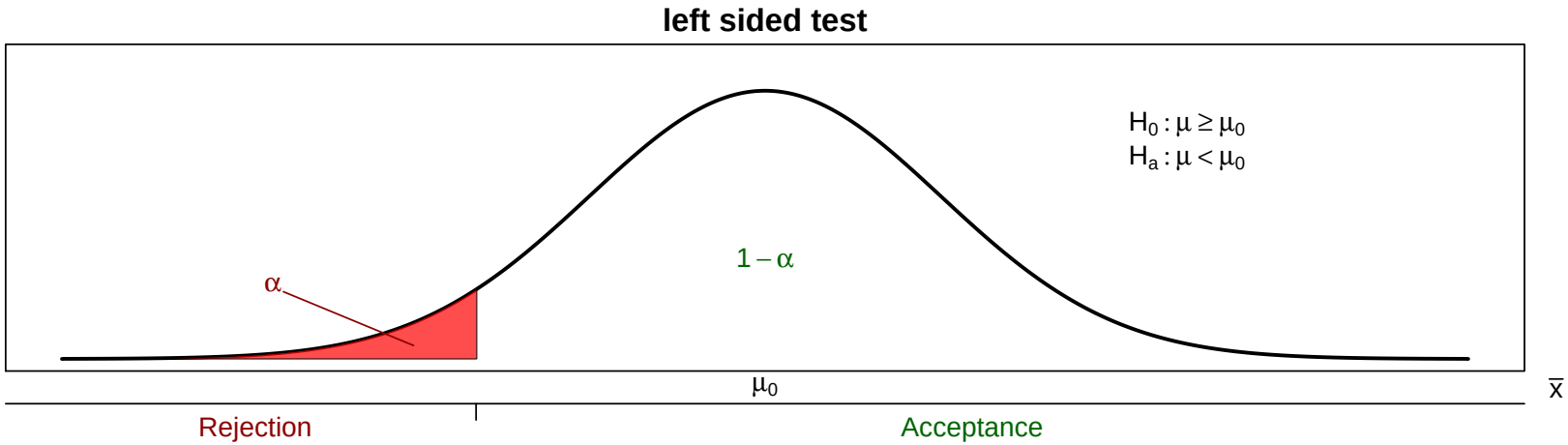
The recipe is now as follows. We reject H_0 if the so called t score is greater than the threshold $t_{n-1,1-\alpha/2}$ which depends on our specified error tolerance α :

$$t = \frac{|\bar{x} - \mu_0|}{SE} > t_{n-1,1-\alpha/2}$$

One sided hypothesis

The situation is somewhat different if we test only for deviations in one direction, i.e. if our Null hypothesis is asymmetric such as

$$H_0 : \mu \geq \mu_0, H_a : \mu < \mu_0 \quad (2)$$



We now reject H_0 if the t score is less than the threshold z_α:

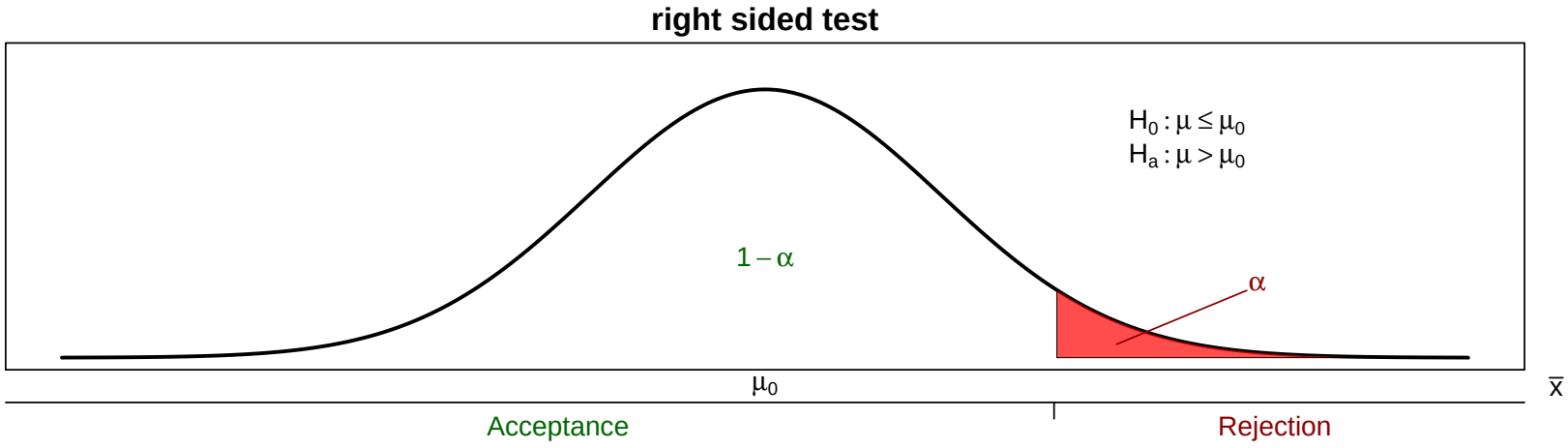
$$t = \frac{\bar{x} - \mu_0}{SE} < z_\alpha$$

Our sample would provide evidence against H_0 only if it is "too small", i.e. if $\bar{x}$ is too far to the left side of μ_0. In that case α is not split among two tails and is the full size of the left tail.

Right tails

And in full analogy for a right sided test:

$$H_0 : \mu \leq \mu_0, H_a : \mu > \mu_0 \quad (2)$$



We now reject H_0 if the t score is greater than the threshold $z_{1-\alpha}$:

$$t = \frac{\bar{x} - \mu_0}{SE} > z_{1-\alpha}$$

Non standard form

The test strategies above are written in the standardized form. Instead we can also rewrite them in a way that compares $\bar{x}$ directly with non standardized thresholds:

- Two-Sided

$$\bar{x} > \mu_0 + z_{1-\alpha/2} \cdot SE \text{ or } \bar{x} < \mu_0 - z_{1-\alpha/2} \cdot SE$$

- Right-sided

$$\bar{x} > \mu_0 + z_{1-\alpha} \cdot SE$$

- Left-sided

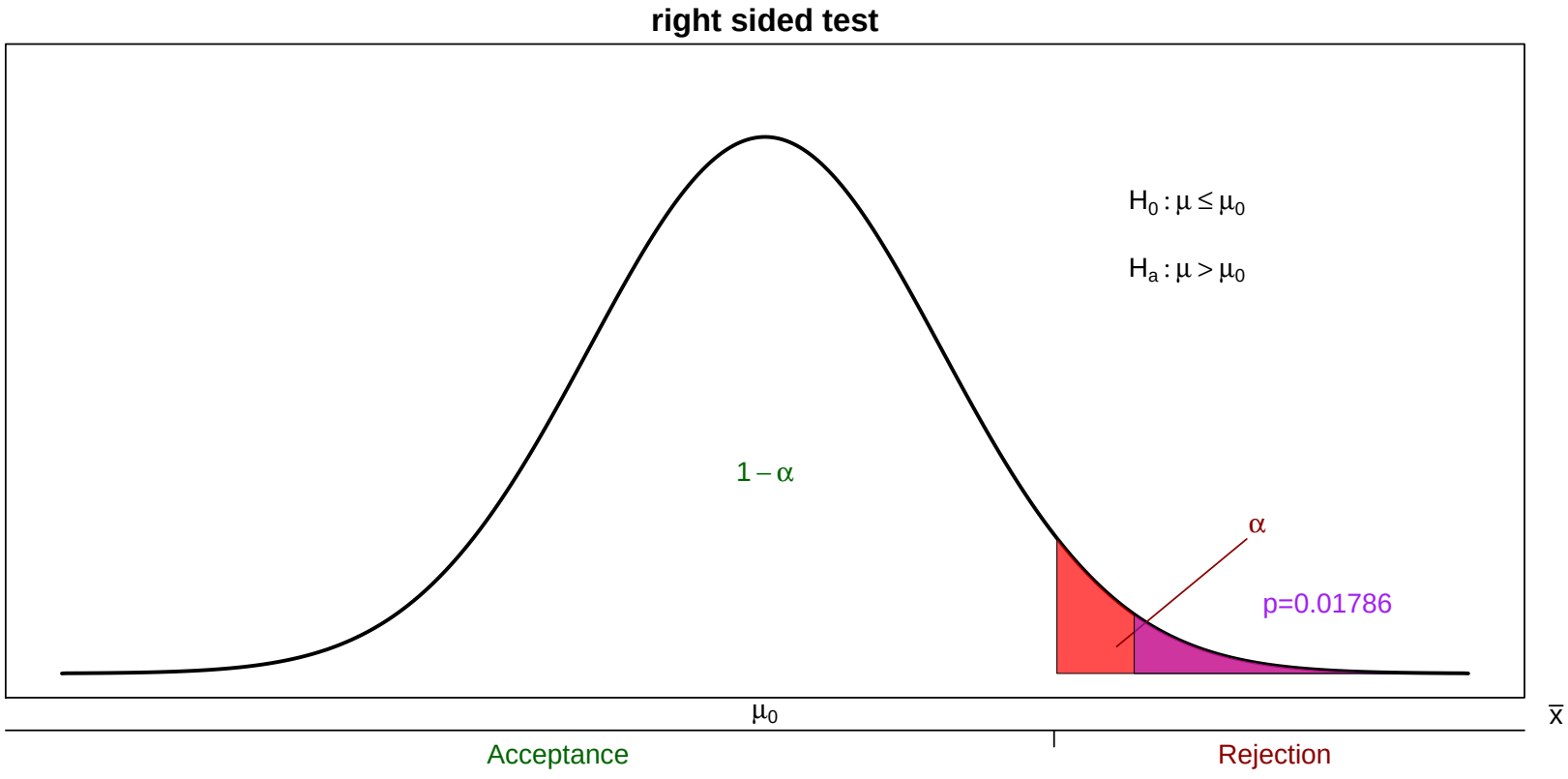
$$\bar{x} < \mu_0 - z_{1-\alpha} \cdot SE$$

p-value

- We use the test statistic to calculate the **p-value**, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is low (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence **reject** H_0.
- If the p-value is high (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence do **not reject** H_0.

p-value, Ex.

For $t = 2.1$ and $\alpha = 0.05$



Interpretation

P Values Are NOT the probability of making a mistake

Incorrect interpretations of P values are very common. The most common mistake is to interpret a P value as the probability of making a mistake by rejecting a true null hypothesis (a Type I error).

There are several reasons why P values can't be the error rate.

First, P values are calculated based on the assumptions that the null is true for the population and that the difference in the sample is caused entirely by random chance. Consequently, P values can't tell you the probability that the null is true or false because it is 100% true from the perspective of the calculations.

Second, while a low P value indicates that your data are unlikely assuming a true null, it can't evaluate which of two competing cases is more likely:

- The null is true but your sample was unusual.
- The null is false.

End of Wooldridge Appendix C

The remaining slides stand on their own, have no analogy in the Wooldridge book

Tschebyschev Inequality

If you do not want to be slave to the strong assumptions of the normal distribution (symmetry, exponential-quadratically decaying tails), you are not left in the complete dark! The following useful inequality holds in the absence of any assumptions:

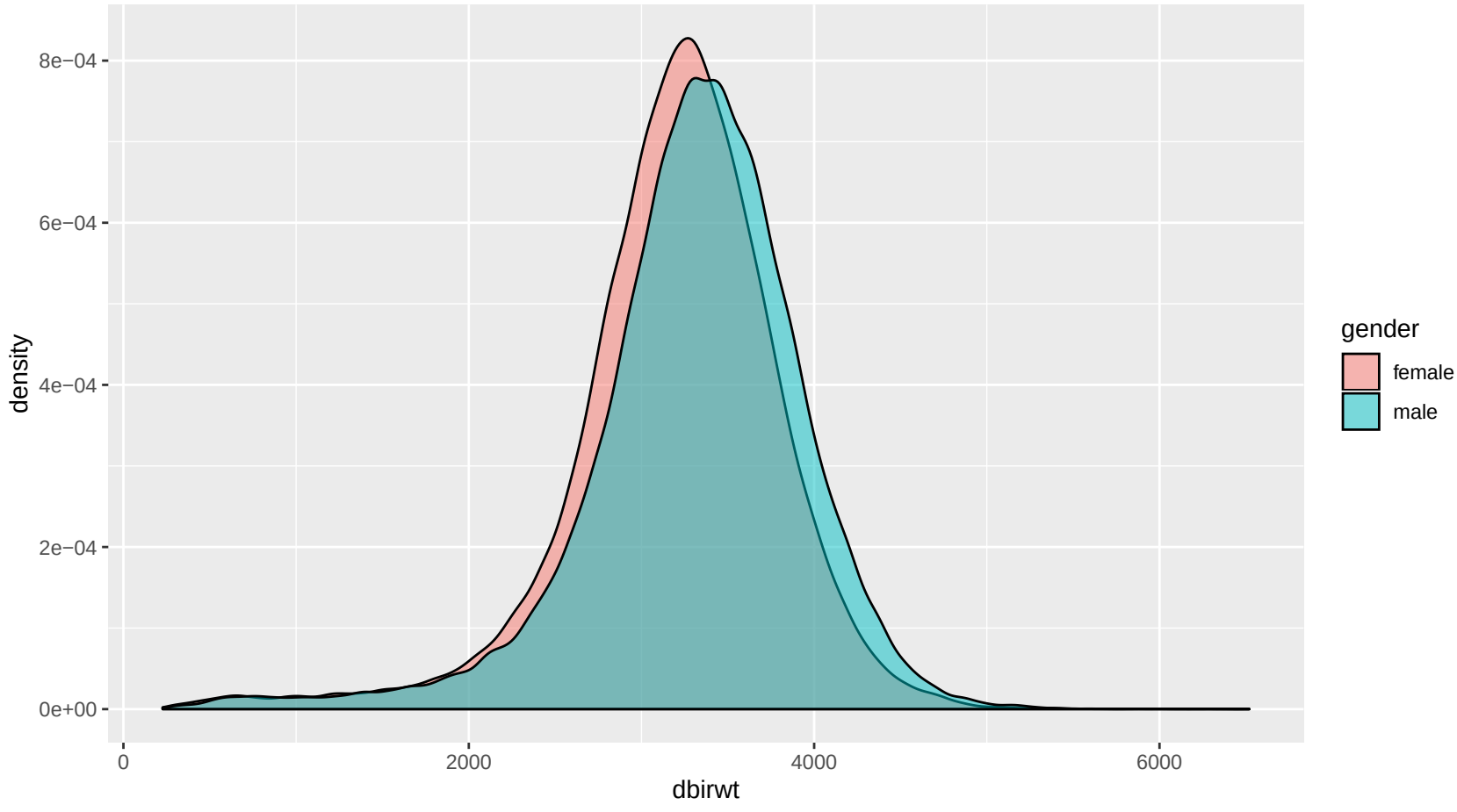
$$P\left(|x-\mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

The differences in the "tail probabilities" are dramatic.

k	Normal	Tschebyschev
2	5 %	25 %
2.5	1 %	16 %
3	0.27 %	11.1 %
4	0.01 %	6.2 %

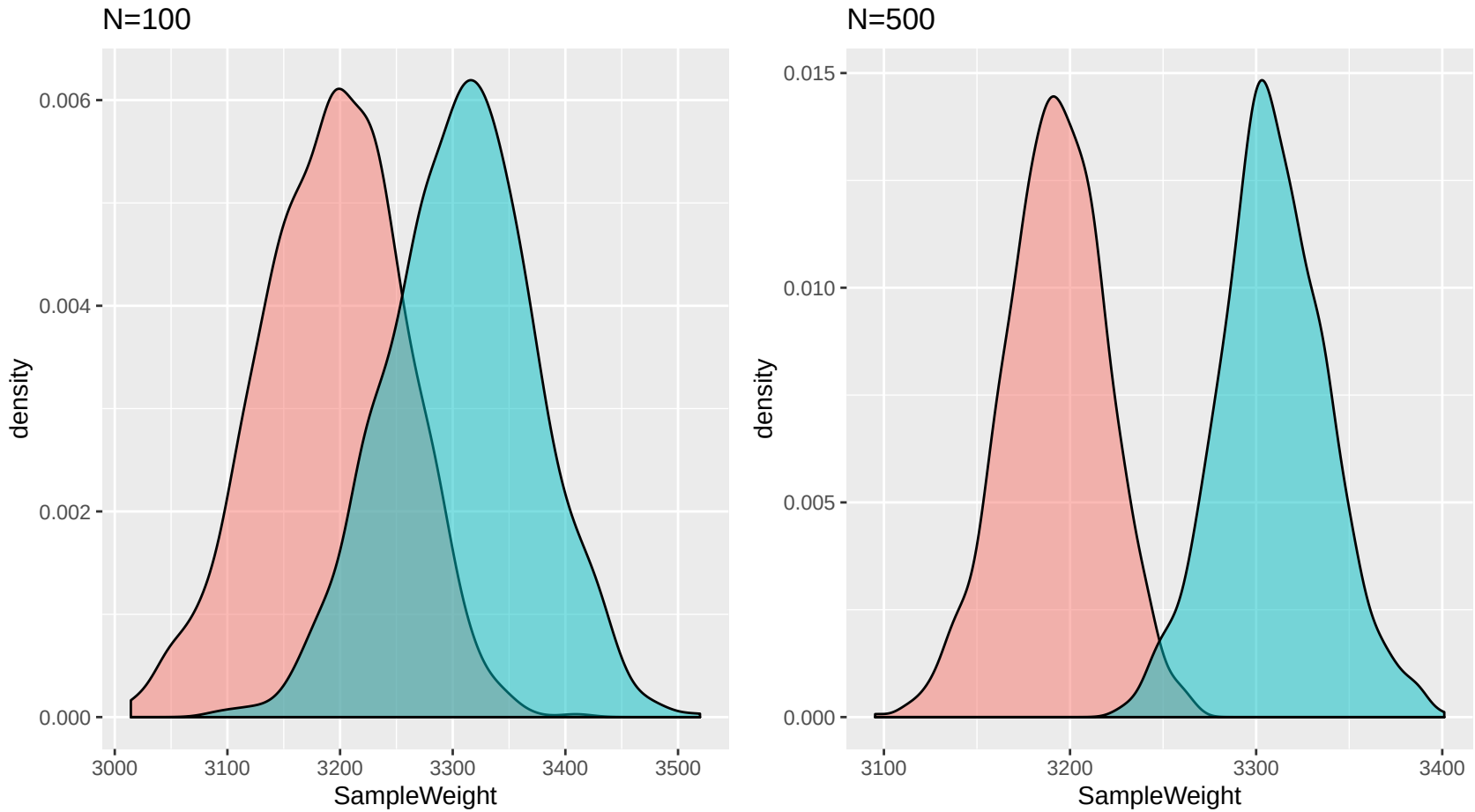
Overlapping Distributions

We have seen that the raw birth data do not allow any conclusions from birth weight to gender as the overlap of the distributions is too high:



Averaging Reduces Variance/Uncertainty

If instead one draws gender specific samples of size e.g. $N = 100$ our "discriminatory power" is greatly increased as the overlap of the distributions shrinks.



Note how much narrower the bell curves become as N grows.

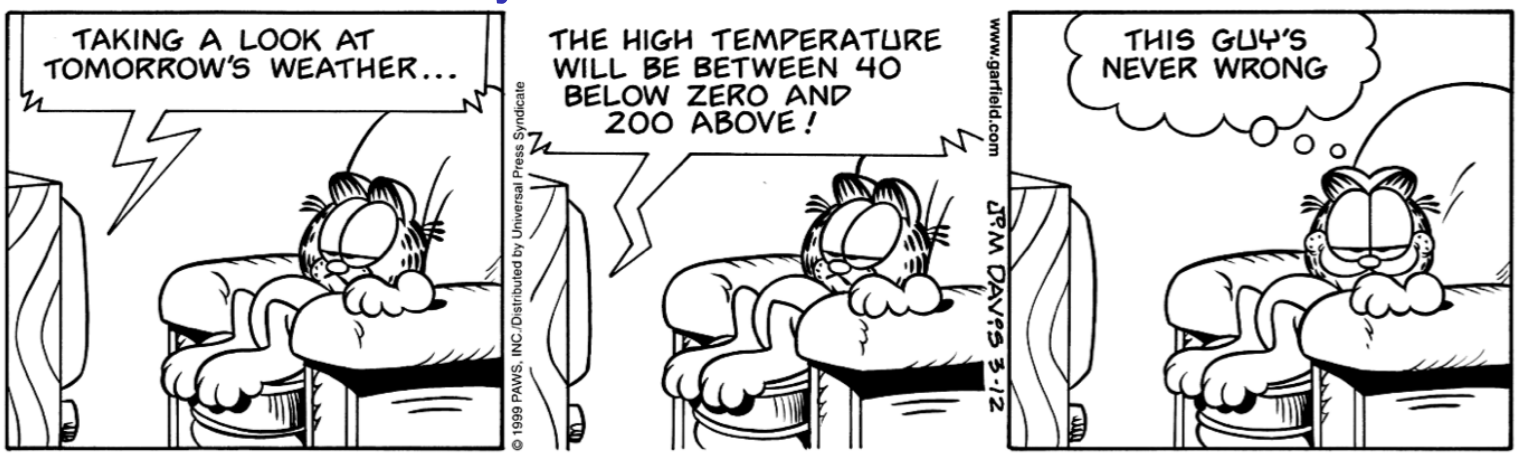
Tasks

The true average weights are $\mu_b = 3310g$ and $\mu_g = 3195g$. Assume that the stdevs are equal for both, $\sigma = 610g$. You take a random sample of $N = 100$ babies of a given sex and find $\bar{x} = 3300g$, $\hat{\sigma} = 600g$.

- For $1 - \alpha = 0.95$ test the hypothesis that this was a sample of baby girls.
- Compute the corresponding p-value.
- Retest using only the p-value.
- Compute the weight threshold beyond which you would reject H_0.
- What is the probability that your decision is wrong?

Power

Isn't a smaller α always better?

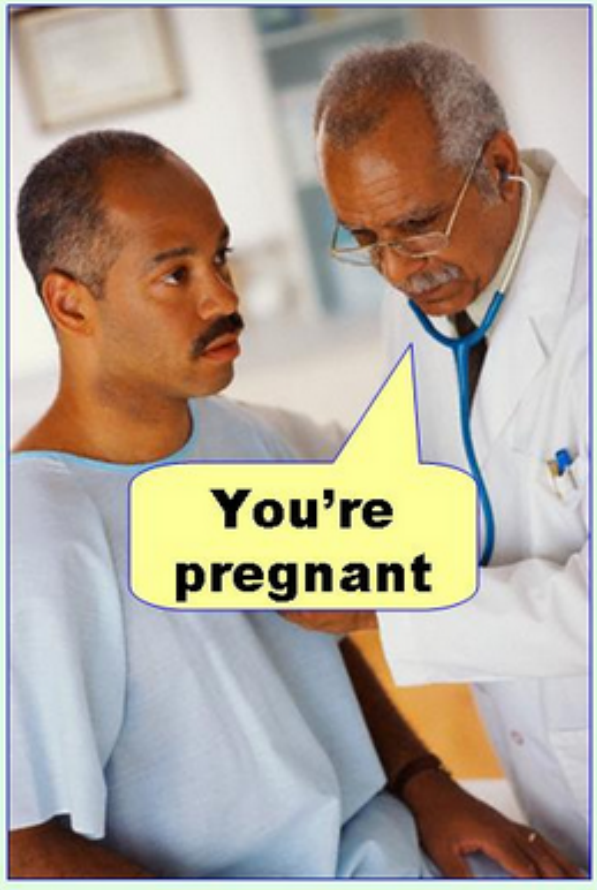


Type I/II errors

Type I/II errors

Type I error

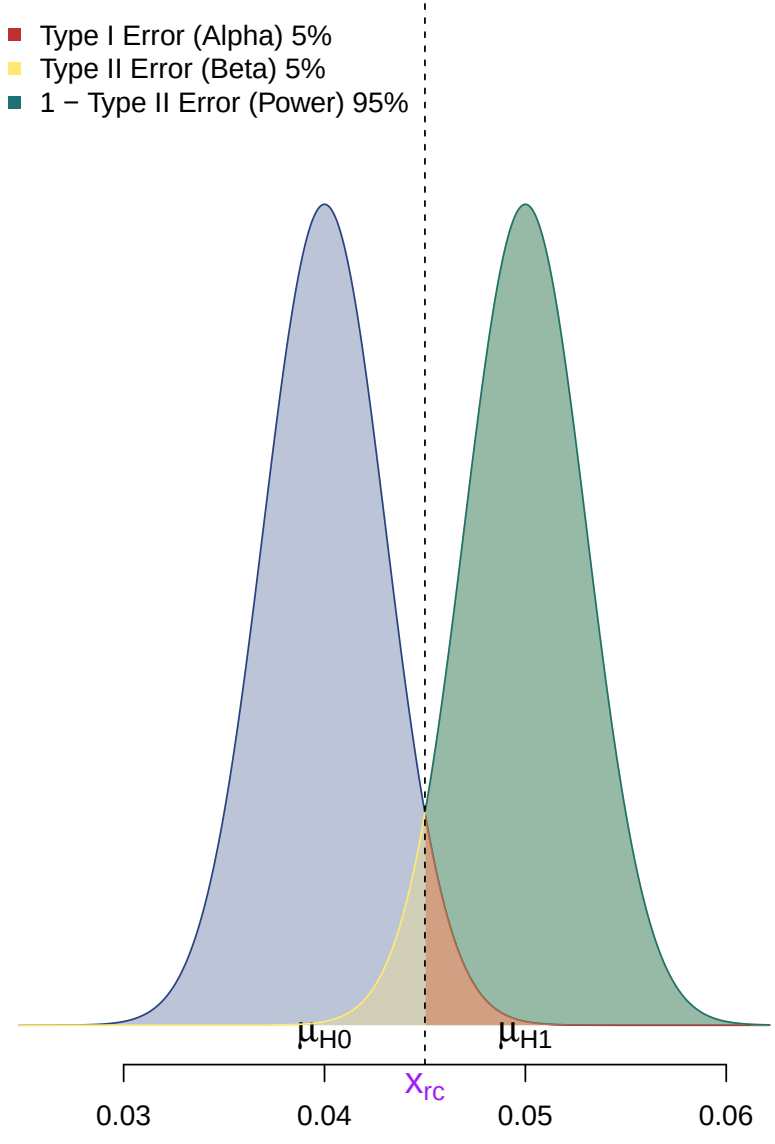
(false positive)

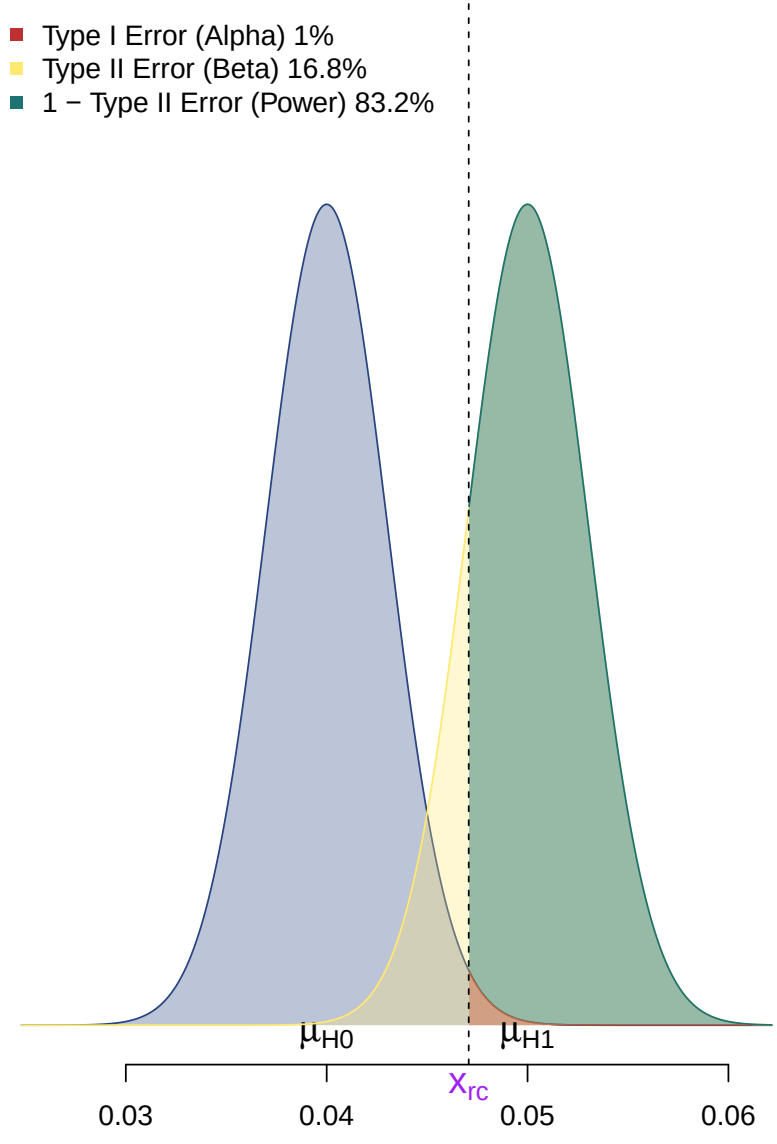


Type II error

(false negative)

Type I/II errors





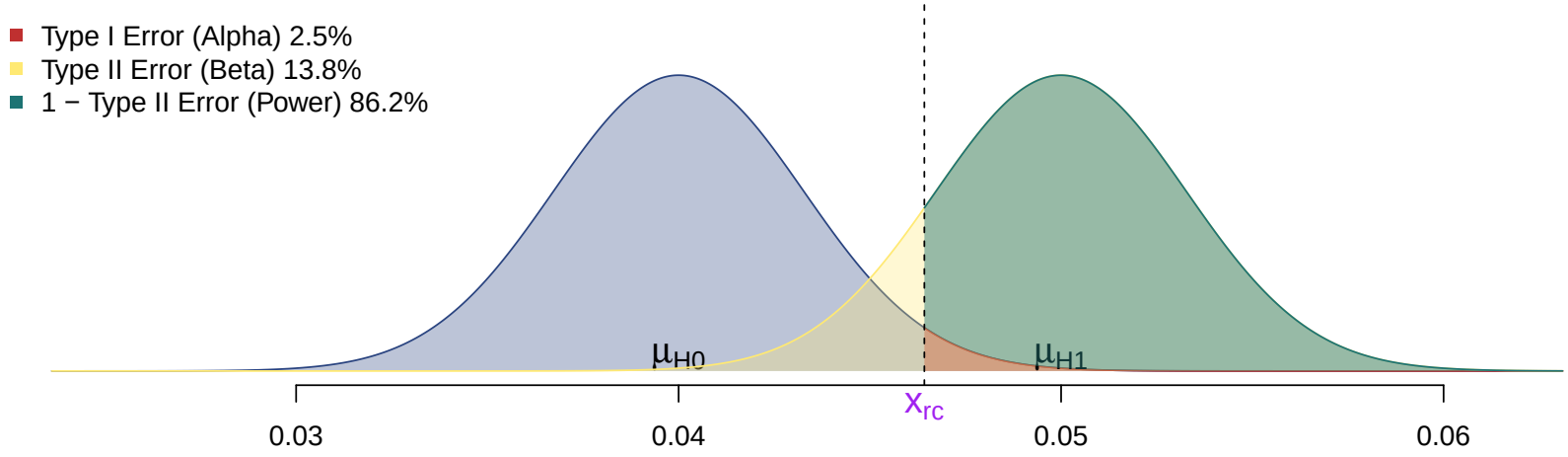
https://www.geogebra.org/m/g5VNUNSU#material/RhuUbJsG

Power

	H_0 true	H_0 false
not reject H_0	OK	type II (β)
reject H_0	type I (α)	OK

- **Type I error** : We reject the Null hypothesis, even though it is true $P(\text{reject } H_0 | H_0)$
- **Type II error** : We do not reject the Null hypothesis, even though it is false $P(\text{not reject } H_0 | \bar{H}_0)$
- Power $= 1 - \beta$

Recipe



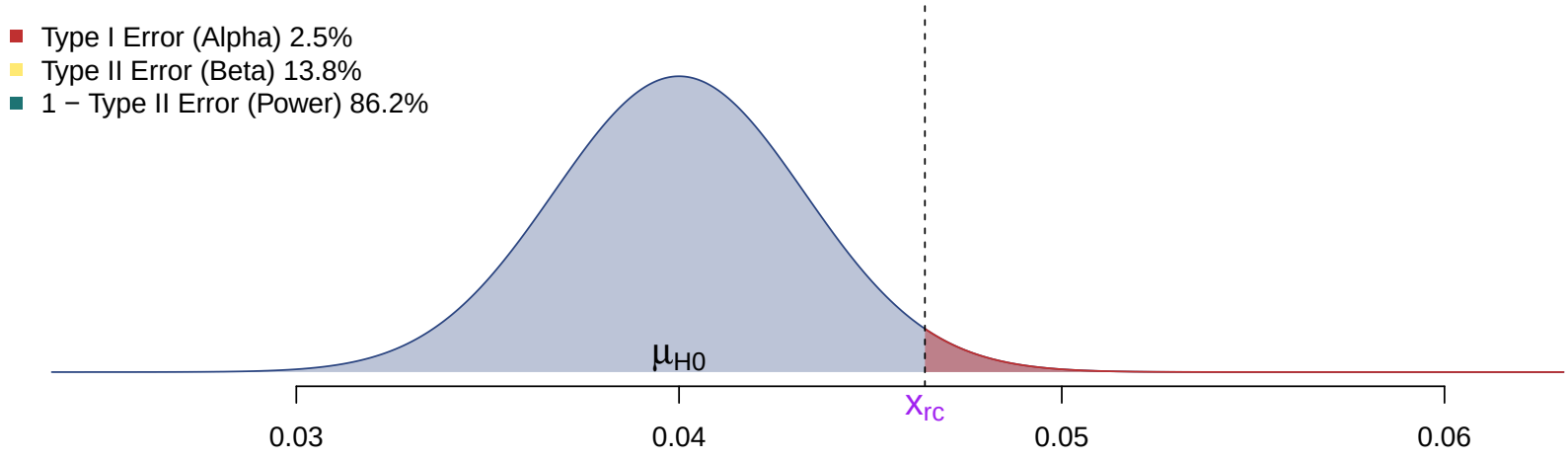
The **right tail** and hence the cutoff are implicitly given by α:

$$x_{rc} = \mu_{H0} + z_{1-\alpha} \cdot SE_0$$

Here, type II error is a **left tail** problem:

$$\beta = P(x \leq x_{rc}|H_1) = P\left(z \leq \frac{x_{rc} - \mu_{H1}}{SE_1}\right)$$

Recipe, Stage I



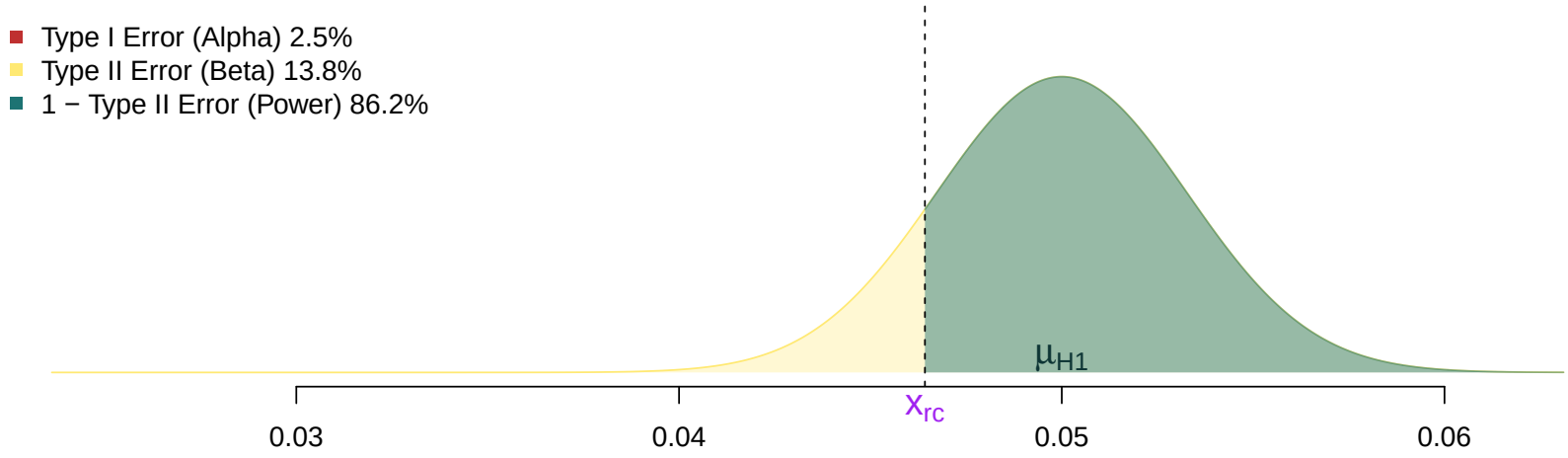
In stage 1 we assume H_0 is true

We use no information from H_1 at all!

The **right tail** and hence the cutoff are implicitly given by α:

$$x_{rc} = \mu_{H0} + z_{1-\alpha} \cdot SE_0$$

Recipe, Stage II



In stage 2 we assume H_1 is true

We use no information from H_0 at all!

Here, type II error is a **left tail** problem:

$$\beta = P(x \leq x_{rc}|H_1) = P\left(z \leq \frac{x_{rc} - \mu_{H1}}{SE_1}\right)$$

Continuous Scores -> binary decision

Medicine

1. PSA "In the past, most doctors considered PSA levels of 4.0 ng/mL and lower as normal. Therefore, if a man had a PSA level above 4.0 ng/mL, doctors would often recommend a prostate biopsy to determine whether prostate cancer was present.However, more recent studies have shown that some men with PSA levels below 4.0 ng/mL have prostate cancer and that many men with higher levels do not have prostate cancer"
2. Bone Density "Normal is a T-score of -1.0or higher. Osteopenia is defined as between -1.0 and -2.5. Osteoporosis is defined as -2.5 or lower, meaning a bone density that is two and a half standard deviations below the mean of a 30-year-old man/woman."

Balancing Act II

Marketing

1. A/B Testing Suppose you've got a conversion rate of 4% on your site. You experiment with a new version of the site that actually generates conversions 5% of the time. You don't know the true conversion rates of course, which is why you're experimenting, but let's suppose you'd like your experiment to be able to detect a 5% conversion rate as statistically significant with 95% probability.
2. In an online experiment you observe 180 clicks out of 4000 impressions.

Balancing Act III

Behavioral Economics

1. "There is no gender difference in lying under stress"

Public Policy

1. Driving Age 16/17/18
2. Global Warming

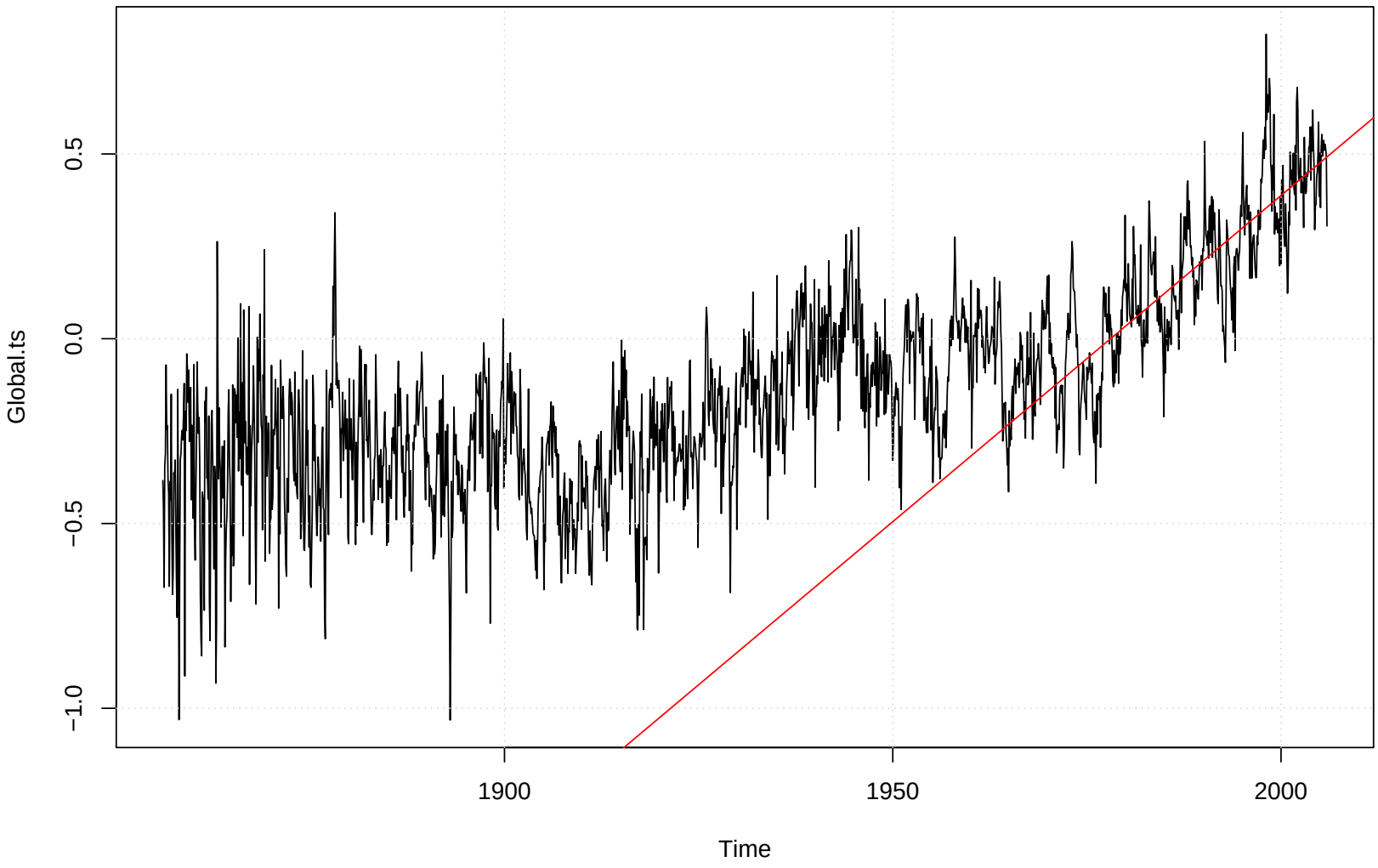
General

1. Car Alarms
2. Airport metal detectors

Table we need to understand

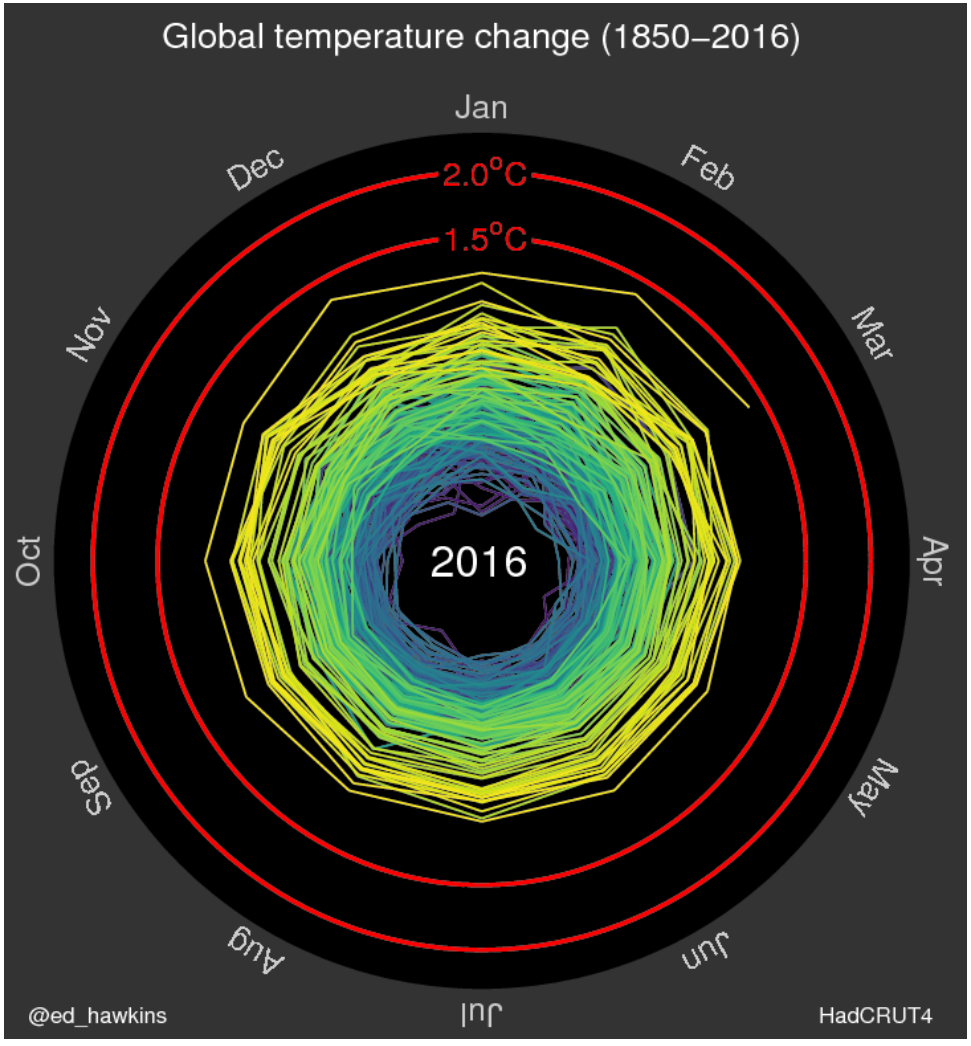
$\bar{x}$	s	n	SE	t	P $\lvert z \rvert > t$

Tasks, power

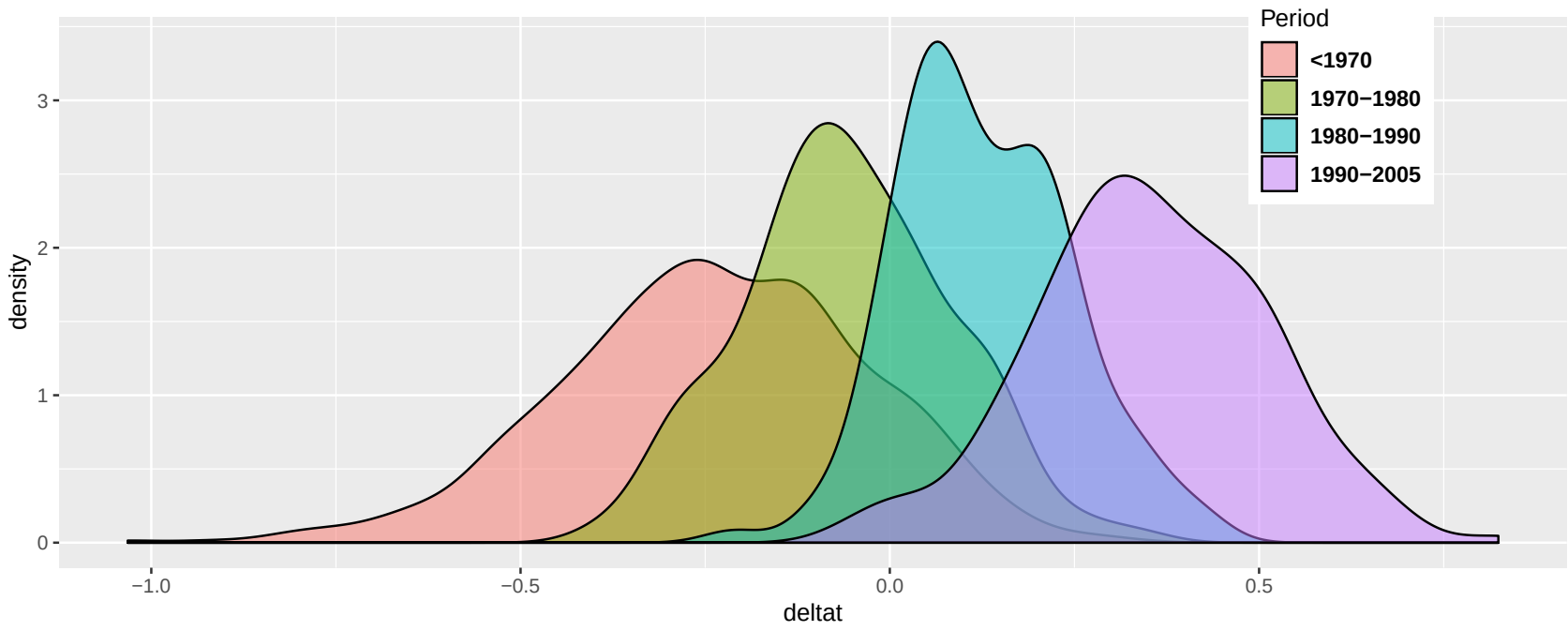


Spiralling global temperatures

Different View at http://www.climate-lab-book.ac.uk/spirals/



Inherent Decision Tradeoff



period	$\bar{x}$	s	n
$<$1970	-0.24	0.20	1369
1970-1980	-0.06	0.14	120
1980-1990	0.13	0.12	120
1990-2005	0.35	0.16	191

General Testing Strategy

- The alternative hypothesis H_A should agree with your evidence/observations!!
- Hence the NULL H_0 hypothesis should contradict your evidence/observations!!
- Compute measure of interest (e.g. sample mean, stdev. etc.) from the sample.
- Compute a **test statistic**, e.g. $z = \sqrt{n} \cdot (\bar{x} - \mu)/s$
- Compare the value of the test statistic to a known distribution
- Obtain a p-value

Sums/differences of RVs

$$X_{1,2} \sim N(\mu_{1,2}, \sigma_{1,2})$$

Two Means

Two samples, two sample means $\bar{x}_{1,2}$, two sample stdevs $s_{1,2}$, two sample sizes $n_{1,2}$!

$$H_0 : \mu_1 = \mu_2, \; H_A : \mu_1 \neq \mu_2$$

- Additional uncertainty !
- For now assume that the true variances are equal: $\sigma_1 = \sigma_2 = \sigma$
- Define $\Delta = \bar{x}_1 - \bar{x}_2$ and

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and

$$\hat{\sigma}_\Delta = \sqrt{\hat{\sigma}^2/n_1 + \hat{\sigma}^2/n_2} = \hat{\sigma}\sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}$$

two-sample t test

https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm

Reject H_0 if

$$\frac{|\Delta|}{\mathrm{SE}_\Delta} = \frac{|\bar{x}_1 - \bar{x}_2|}{\hat{\sigma}_\Delta} > t_{n_1+n_2-2,1-\alpha/2}$$

Tasks:

- Write down the equivalent test decisions for one-sided hypotheses.
- Test equality of means for the temperature periods.

ChiSquare

Squaring Normals

The chi-squared (χ^2_n) distribution with n degrees of freedom is the distribution of the **sum of the squares of n independent standard normal** random variables

$$Z \overset{iid}{\sim} N(0;1) \Rightarrow \sum_{i=1}^{n} Z^2 \sim \chi^2_n$$

$$Y \overset{iid}{\sim} N(\mu;\sigma) \Rightarrow \sum_{i=1}^{n} [\qquad\qquad\qquad\qquad] = \sum_{i=1}^{n} [\qquad\qquad\qquad\qquad] \sim \chi^2_n$$

$$Y \overset{iid}{\sim} N(\mu;\sigma) \Rightarrow \sum_{i=1}^{n} (Y_i - \bar{Y})^2/\sigma^2 \sim \chi^2_?$$

$$Y \overset{iid}{\sim} N(\mu;\sigma) \Rightarrow \hat{\sigma}^2 \sim$$

Squaring Normals

$$Z \overset{iid}{\sim} N(0;1) \Rightarrow \sum_{i=1}^{n} Z^2 \sim \chi^2_n$$

$$Y \overset{iid}{\sim} N(\mu;\sigma) \Rightarrow \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma^2}\right)^2 = \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2} \sim \chi^2_n$$

$$Y \overset{iid}{\sim} N(\mu;\sigma) \Rightarrow \sum_{i=1}^{n} (Y_i - \bar{Y})^2/\sigma^2 = SS_Y/\sigma^2 \sim \chi^2_{n-1}$$

$$Y \overset{iid}{\sim} N(\mu;\sigma) \Rightarrow (n-1) \cdot \frac{SS_Y/(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

Sample Variances follow a chi-square distribution

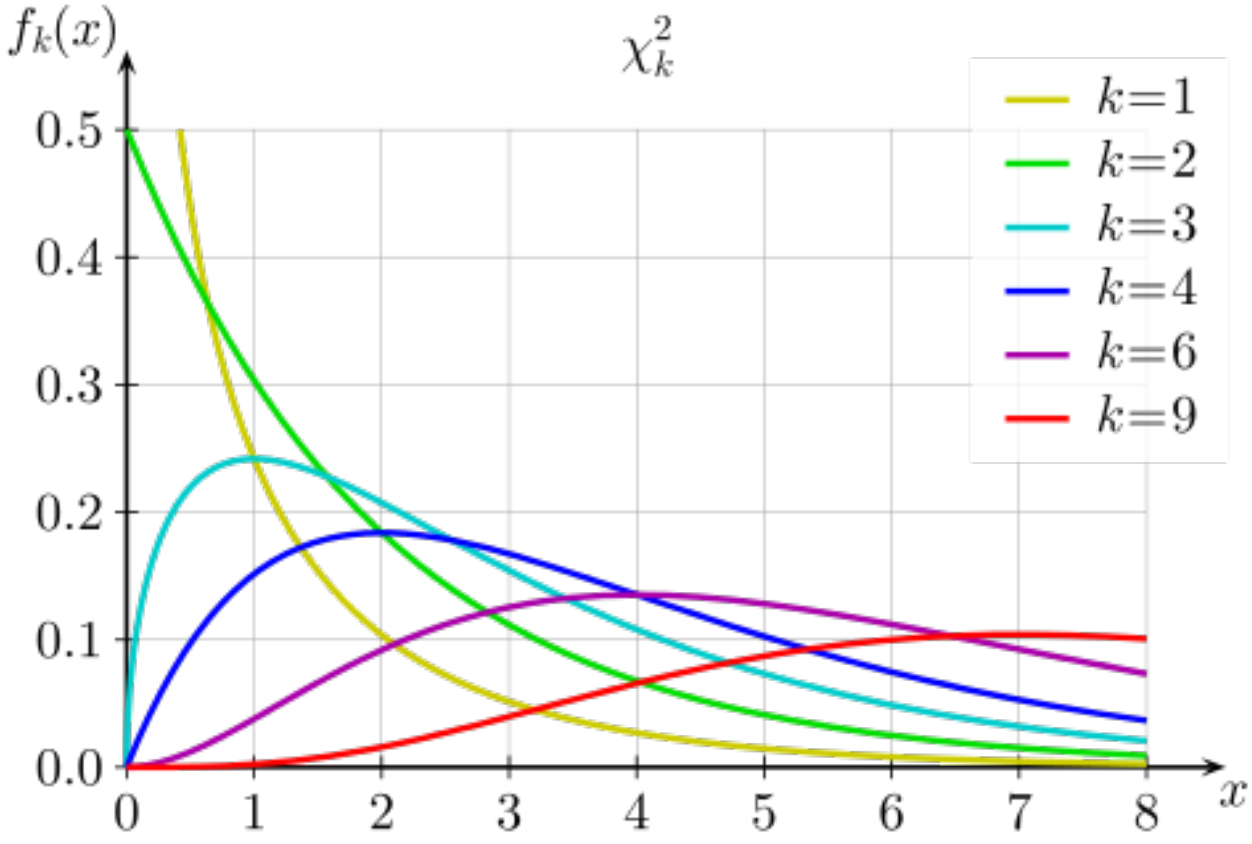
If the data follow a Normal distribution $x \overset{iid}{\sim} N(\mu, \sigma^2)$, the sample variance follows a chi-square distribution with $n-1$ degrees of freedom:

$$(s^2) = \hat{\sigma}^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1} \Leftrightarrow \frac{(n-1) \cdot \hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

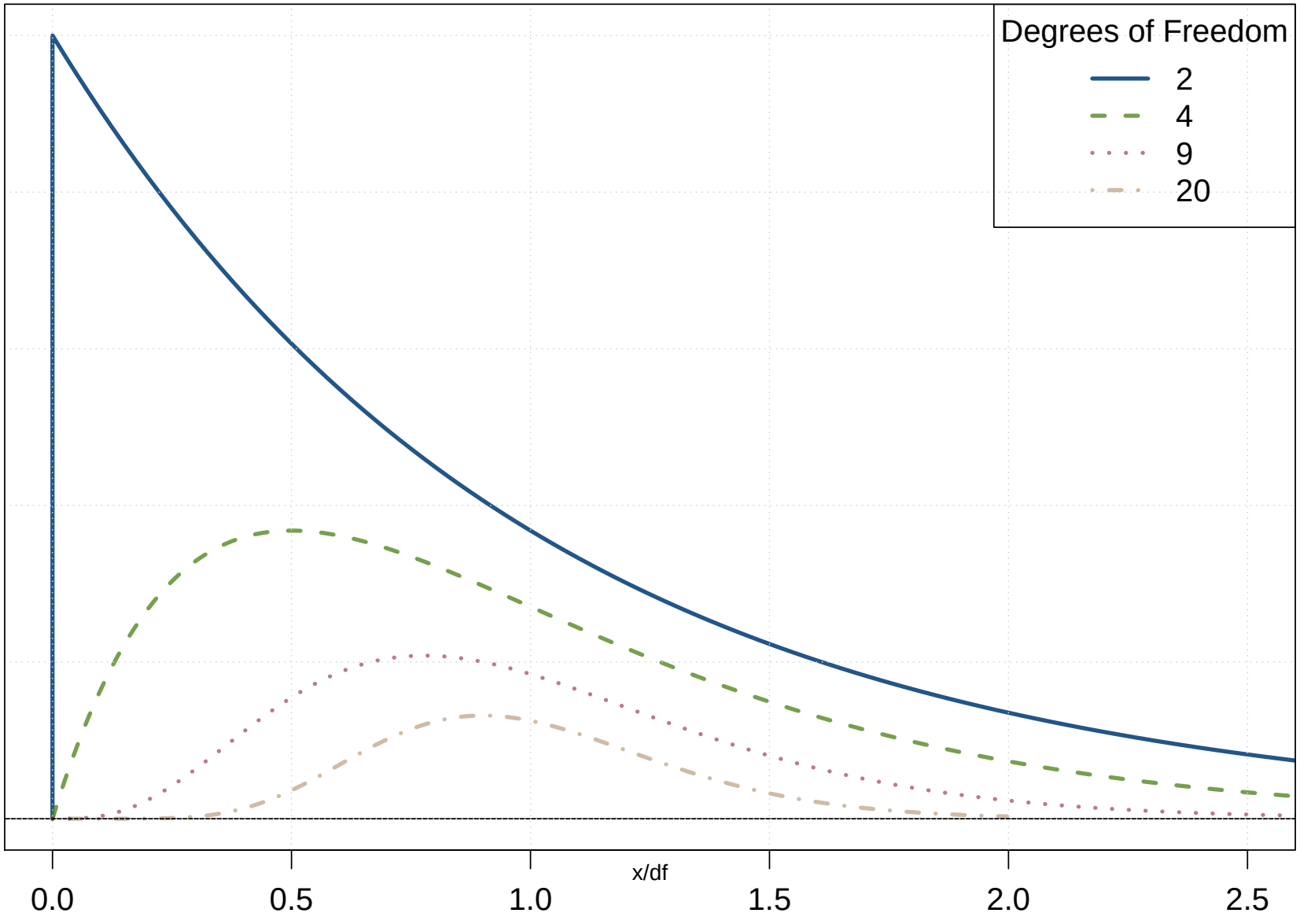
Comment: The t-distribution is "officially" defined as the ratio of a standard normal Z over a scaled version of $U \sim \chi^2_n$

$$t_n \sim \frac{Z}{\sqrt{Y/n}} \Rightarrow \frac{\sqrt{n} \cdot (\bar{x} - \mu)}{\hat{\sigma}} \sim t_{n-1}$$

χ^2 density



χ^2 density, rescaled



Testing Variances

Confidence Interval for variances:

$$\left[\frac{(n-1)\cdot\hat{\sigma}^2}{\chi^2_{(n-1,1-\alpha/2)}};\frac{(n-1)\cdot\hat{\sigma}^2}{\chi^2_{(n-1,\alpha/2)}}\right] \tag{1}$$

Reject Null hypothesis H_0 if

$H_0 : \sigma^2 = \sigma_0^2$, $H_A : \sigma^2 \neq \sigma_0^2$.

$$\frac{(n-1)\cdot\hat{\sigma}^2}{\sigma_0^2} < \chi^2_{(n-1,\alpha/2)}, \text{ or } > \chi^2_{(n-1,1-\alpha/2)} \tag{2}$$

$H_0 : \sigma^2 \geq \sigma_0^2$, $H_A : \sigma^2 < \sigma_0^2$.

$$\frac{(n-1)\cdot\hat{\sigma}^2}{\sigma_0^2} < \chi^2_{(n-1,\alpha)} \tag{3}$$

ChiSquare Table

(row labels are degrees of freedom, column header is right tail α:)

df / α	**0.99**	**0.975**	**0.95**	**0.9**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88	10.83
2	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60	13.82
3	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84	16.27
4	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75	20.52
6	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55	22.46
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28	24.32
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95	26.12
9	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59	27.88
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19	29.59
15	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80	37.70
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00	45.31
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93	52.62
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67	59.70
40	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77	73.40
50	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49	86.66
60	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95	99.61
70	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.21	112.32
80	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32	124.84
90	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	128.30	137.21
100	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17	149.45
125	91.18	95.95	100.18	105.21	145.64	152.09	157.84	164.69	169.47	179.60
150	112.67	117.98	122.69	128.28	172.58	179.58	185.80	193.21	198.36	209.26
175	134.44	140.26	145.41	151.49	199.36	206.87	213.52	221.44	226.94	238.55
200	156.43	162.73	168.28	174.84	226.02	233.99	241.06	249.45	255.26	267.54
225	178.61	185.35	191.28	198.28	252.58	260.99	268.44	277.27	283.39	296.29
250	200.94	208.10	214.39	221.81	279.05	287.88	295.69	304.94	311.35	324.83

Tasks, Variance Testing

- Could it be that the global temperature stdevs 1990-2005 is truly less than 0.2 ?
- Test for different stdevs in boys and girls from the article *Population sex differences in IQ at age 11: the Scottish mental survey 1932*, Intelligence 31 (2003) 533-542.
  This left 79,376 (39,343 girls and 40,033 boys) with total score information. The mean IQ score was 100.64 for girls and 100.48 for boys, the standard deviation was 14.1 for girls and 14.9 for boys.

Scottish mental survey 1932

The mean IQ score, based on the total score of the Picture and Moray House Tests, was 100.64 for girls and 100.48 for boys. This difference of 0.16 (95% CI = – 0.037 to 0.367) was nonsignificant, despite the massive numbers tested, $t(79{,}374) = 1.6$, $P = .11$. The standard deviation was 14.1 for girls and 14.9 for boys. Levene's test for comparing variances was significant, $P < .001$, i.e., boys and girls differed significantly in variability.

Participants were allocated to IQ score bands as described above. The absolute number and the relative percentages of boys and girls in each IQ band is shown in Fig. 1. The proportions

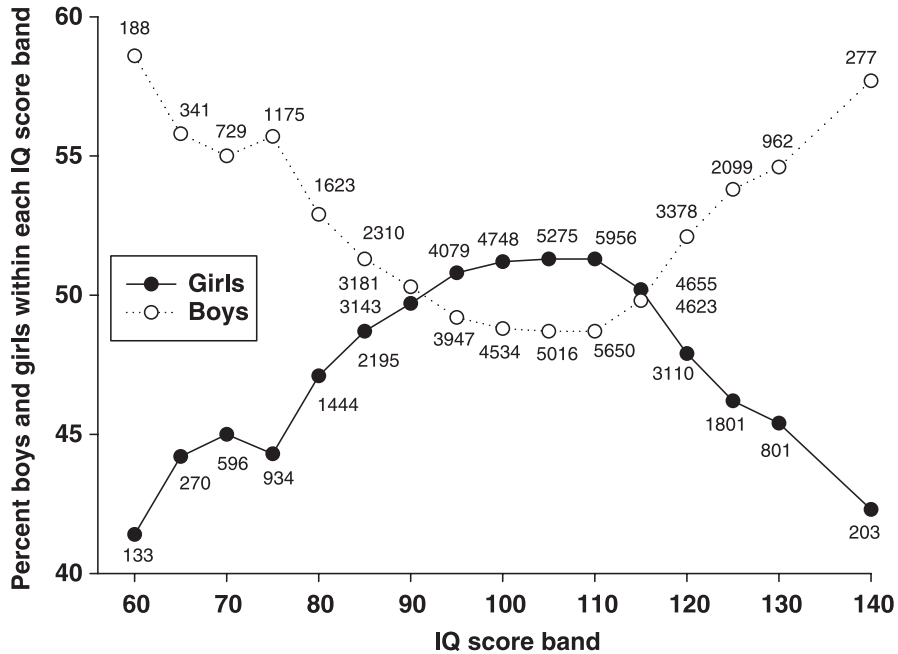


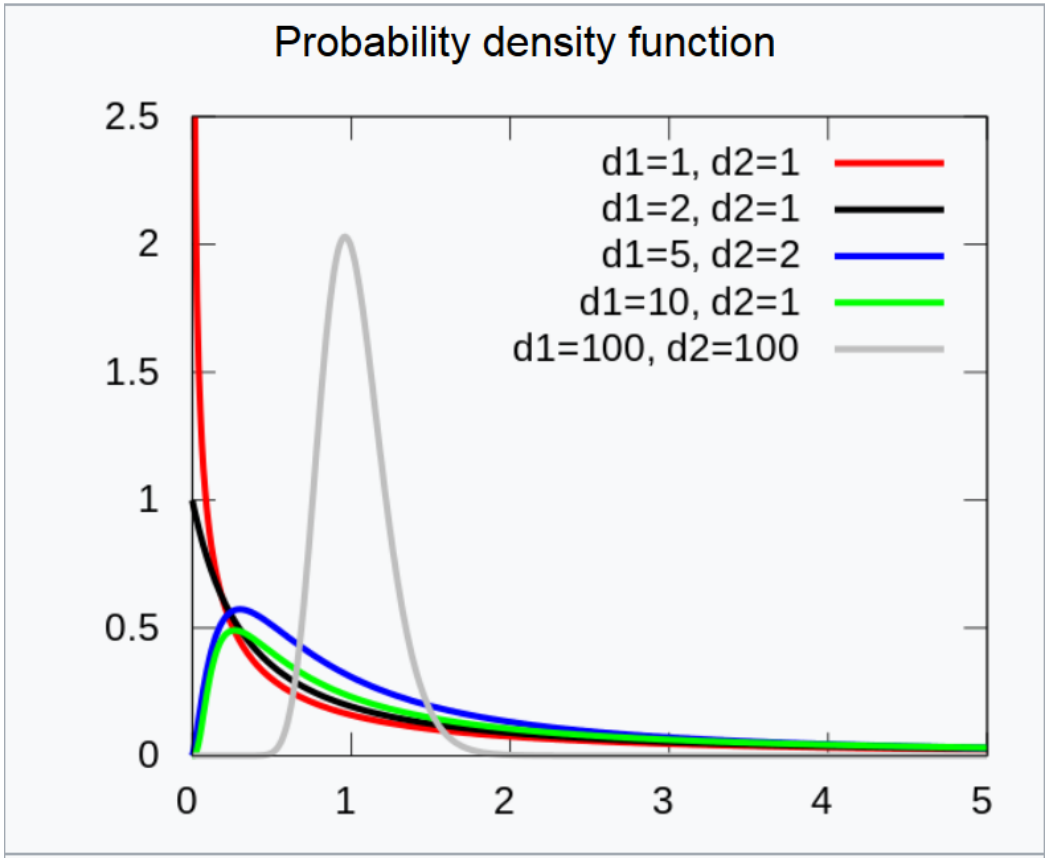
Fig. 1. Numbers and percentages of boys and girls found within each IQ score band of the Scottish population born in 1921 and tested in the Scottish Mental Survey in 1932 at age 11. The *y* axis represents the percentage of each sex in each 5-point band of IQ scores. Numbers beside each point represent the absolute numbers of boys and girls in each 5-point IQ score band.

F Test

The F distribution

It is the ratio of two appropriately scaled chi-squared variables $U_i \sim \chi^2_{d_i}$:

$$F \sim \frac{U_1/d_1}{U_2/d_2}$$



Two Variances

$$
\begin{aligned}
H_0 &: \quad \sigma_1^2 = \quad \sigma_2^2 \\
H_a &: \quad \sigma_1^2 < \quad \sigma_2^2 \text{ for a lower one-tailed test} \\
H_a &: \quad \sigma_1^2 > \quad \sigma_2^2 \text{ for an upper one-tailed test} \\
H_a &: \quad \sigma_1^2 \neq \quad \sigma_2^2 \text{ for a two-tailed test}
\end{aligned}
$$

Test Statistic

$$F = \frac{s_1^2}{s_2^2}$$

The hypothesis that the two variances are equal is rejected if

$$
\begin{aligned}
F &< F_{\alpha,n_1-1,n_2-1} \text{ for a lower one-tailed test} \\
F &> F_{\alpha,n_1-1,n_2-1} \text{ for an upper one-tailed test} \\
F &< F_{\alpha/2,n_1-1,n_2-1} \text{ or } F > F_{1-\alpha/2,n_1-1,n_2-1} \text{ for a two-tailed test}
\end{aligned}
$$

Exercise, 2 variances

NOTE: The above follows the convention that $F_{1-\alpha}$ is the **upper** critical value from the F distribution and F_{α} is the **lower** critical value from the F distribution. Note that this is the opposite of the designation used by some texts and software programs.

Task: Test if temperature fluctuations >1970 are about the same as < 1970

period	$\bar{x}$	s	n
<1970	-0.24	0.20	1369
1970-2005	0.18	0.22	431

Shortcomings of classical Tests

Bayesian Critique I

While classical hypothesis testing has a long and celebrated history in the statistical literature and continues to be a favorite of practitioners, several fairly substantial criticisms of it may be made. First, the approach can be applied straightforwardly only when the two hypotheses in question are nested, one within the other. That is, H_0 must be a simplification of H_a (say, by setting one of the model parameters in H_a equal to a constant – usually zero). But many practical hypothesis testing problems involve a choice between two (or more) models that are not nested (e.g., choosing between quadratic and exponential growth models, or between exponential and lognormal error distributions).

Bayesian Critique II

A second difficulty is that tests of this type can only offer evidence against the null hypothesis. A small p-value indicates that the larger, alternative model has significantly more explanatory power. However, a large p-value does not suggest that the two models are equivalent, but only that we lack evidence that they are not. This difficulty is often swept under the rug in introductory statistics courses, the technically correct phrase "fail to reject [the null hypothesis]" being replaced by "accept."

Bayesian Critique III

Third, the p-value itself offers no direct interpretation as a "weight of evidence," but only as a long-term probability (in a hypothetical repetition of the same experiment) of obtaining data at least as unusual as what was actually observed. Unfortunately, the fact that small p-values imply rejection of H_0 causes many consumers of statistical analyses to assume that the p-value is "the probability that H_0 is true," even though its definition falls short of this sweeping conclusion. A final, somewhat more philosophical criticism of p-values is that they depend not only on the observed data, but also the total sampling probability of certain unobserved datapoints; namely, the "more extreme" values of the test statistic.. As a result, two experiments with identical likelihoods could result in different p-values if the two experiments were designed differently.

The Likelihood Principle

In conjunction with the Example below, this fact violates a proposition known as the Likelihood Principle (Birnbaum, 1962), which can be stated briefly as follows:

In making inferences or decisions about θ after y is observed, all relevant experimental information is contained in the likelihood function for the observed y.

By taking into account not only the observed data y, but also the unobserved but more extreme values of Y, classical hypothesis testing violates the Likelihood Principle.

Example by Lindley and Phillips (1976)

Suppose in 12 independent tosses of a coin, I observe 9 heads and 3 tails. I wish to test the null hypothesis $H_0 : \theta = 1/2$ versus the alternative hypothesis $H_a : \theta > 1/2$, where θ is the true probability of heads. Given only this much information, two choices for the sampling distribution emerge:

1. *Binomial*: The number $n = 12$ tosses was fixed beforehand, and the random quantity X was the number of heads observed in the n tosses. Then $X \sim \text{Bin}(12, \theta)$, and the likelihood function is given by

$$L_1(\theta) = \binom{12}{9} \theta^9 \cdot (1-\theta)^3$$

2. *Negative Binomial*: Data collection involved flipping the coin until the third tail appeared. Here, the random quantity X is the number of heads required to complete the experiment, so that $X \sim \text{NegBin}(r = 3, \theta)$

$$L_2(\theta) = \binom{r+x-1}{x} \theta^x \cdot (1-\theta)^r = \binom{11}{9} \theta^9 \cdot (1-\theta)^3$$

Example, cont.

We can compute the p-value for both examples:

1.

$$p_1 = P_{\theta=0.5}(X \geq 9) = \sum_{j=9}^{12} \binom{12}{j} \theta^j \cdot (1-\theta)^{12-j} = 0.075$$

2.

$$p_2 = P_{\theta=0.5}(X \geq 9) = \sum_{j=9}^{\infty} \binom{2+j}{j} \theta^j \cdot (1-\theta)^3 = 0.075$$

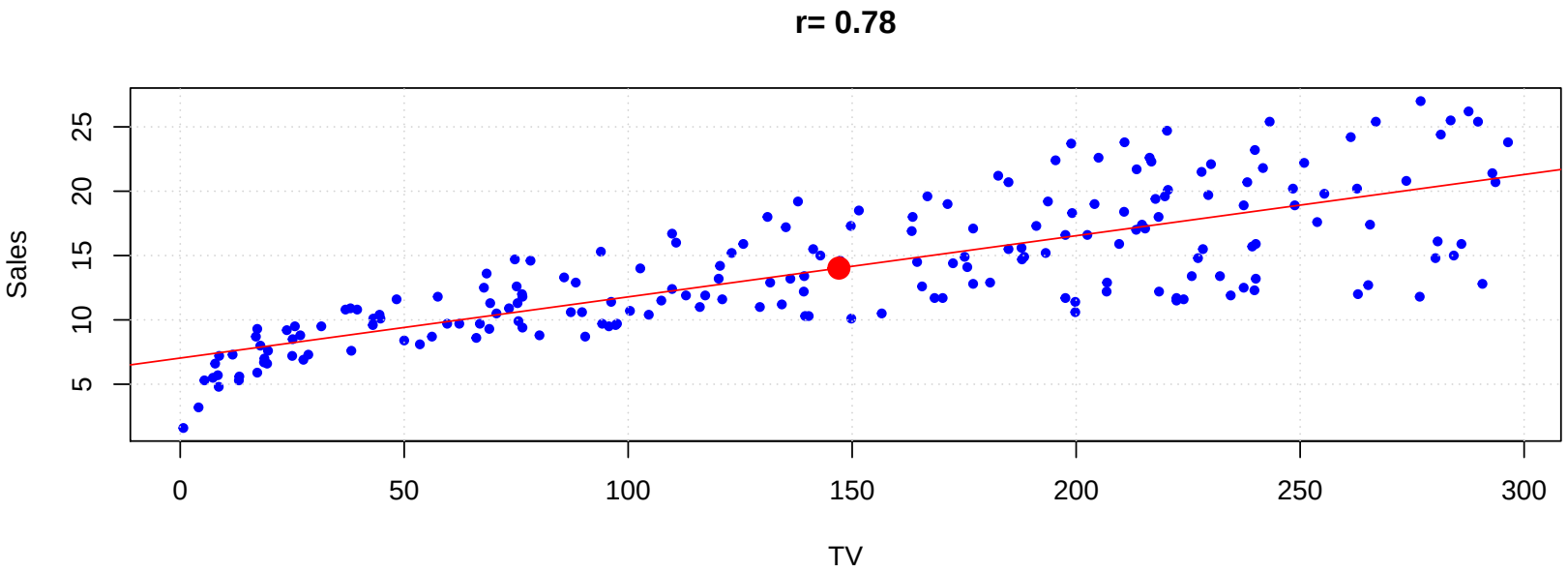
Thus, using the "usual" Type I error level $\alpha = .05$, we see that the two model assumptions lead to two different decisions: we would reject H_0 if X were assumed negative binomial, but not if it were assumed binomial.

Appendix

Sums of random variables

"Variances add"

Heteroskedasticity



1. Test for the right half of the data to agree with the estimated $\hat{\sigma}_u = 3.26$ ($N = 200$)
2. Test for equal variances "left/right half".