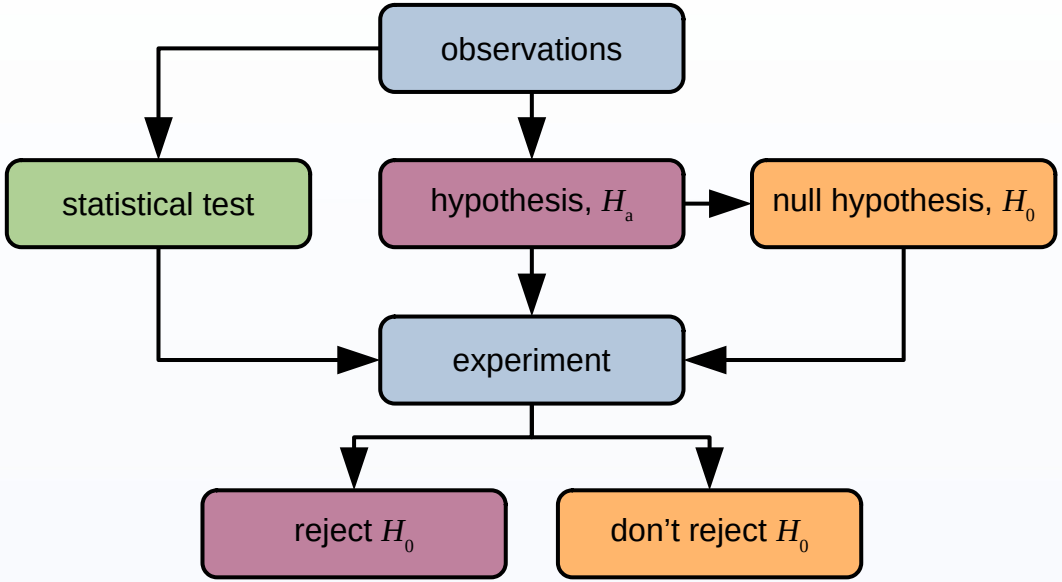
EGM101 – Skills Toolbox

Week 8, Part 2: Hypothesis Testing

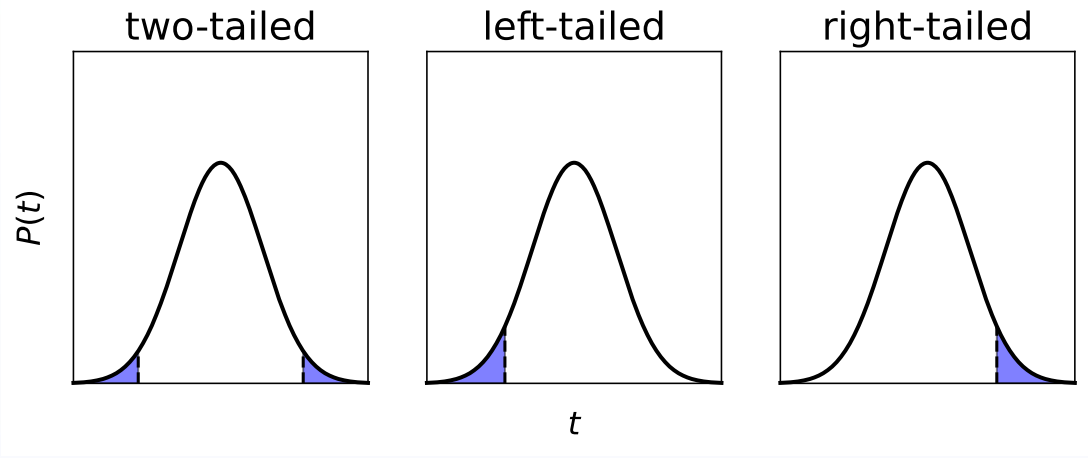
Inferential statistics

- Recall (W5, P1):
  - Descriptive statistics: organizing and summarizing data
  - Inferential statistics: drawing conclusions from "good" data
- Formal process of drawing conclusions from "good" data: hypothesis testing
- The null hypothesis, H_0, is the contradiction (logical negation) of the alternative hypothesis, H_a
  - e.g., "treatment made no difference"
- Note: we cannot prove that the alternative hypothesis is true!
- Goal: falsify (refute) alternative hypothesis by attempting to refute the null hypothesis



Choose your hypothesis

- First off: testing population mean, μ
  - "One sample" test
  - Later: testing sample means, etc.
- H_0 usually something like:
  - H_0: $\mu = \mu_0$
- Two-tailed test:
  - H_a: $\mu \neq \mu_0$
- One-tailed test:
  - H_a: $\mu < \mu_0$ (left-tailed)
  - H_a: $\mu > \mu_0$ (right-tailed)

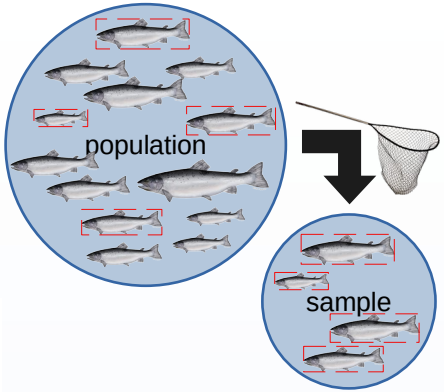


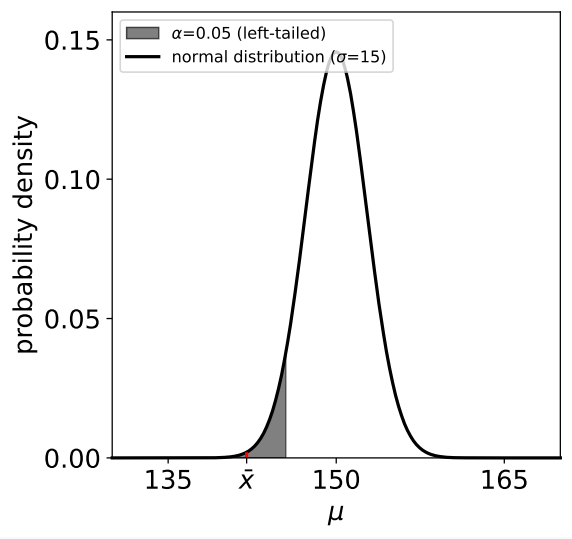
What are we "testing," anyway?

- Example: salmon length
- Hypotheses:
  - H_0: $\mu = 150$ cm
  - H_a: $\mu < 150$ cm
- Question: what is the probability of seeing observations *at least this extreme*, assuming that H_0 is correct? $\rightarrow P(E|H_0)$
- Z-test: compare z-score of sample mean
  - Test statistic, Z: z-score using $\sigma_{\bar{x}}$
- What is probability of getting a z-score lower than -2.921?
  - $p = 0.002$
- Very low $\rightarrow \mu$ is most likely less than 150 cm
  - Formally: "we reject H_0 in favor of H_a"

population: $\mu = 150\,\text{cm}$, $\sigma = 15\,\text{cm}$

sample ($n = 30$): $\bar{x} = 142\,\text{cm}$, $s = 20\,\text{cm}$





$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$= \frac{142 - 150}{15/\sqrt{30}}$$

$$= -2.921$$

Type I & Type II Errors, revisited

- For hypothesis testing:
  - Type I Error/False Positive: incorrectly reject H_0
  - Type II Error/False Negative: incorrectly fail to reject H_0
- Probabilities α, β:
  - What we use to make our decision
  - Outcome of test should be less likely than a false positive (incorrectly rejecting H_0)
  - Power of the test $(1-\beta)$: probability of a true positive (correctly rejecting H_0)

Decision	H_0 is: False	H_0 is: True
Reject H_0	Correct (True Positive)	Type I (False Positive)
Do not reject H_0	Type II (False Negative)	Correct (True Negative)

$$P(True\ Positive) = 1-\beta \qquad P(False\ Positive) = \alpha$$

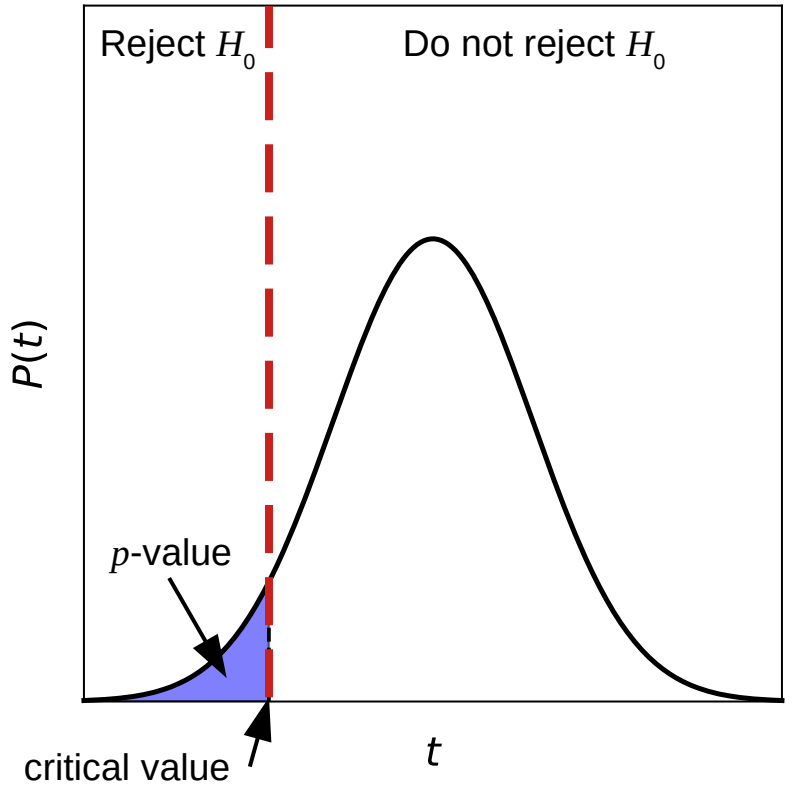
$$P(False\ Negative) = \beta \qquad P(True\ Negative) = 1-\alpha$$

Statistical significance

- In a statistical context, “significant” does not mean:
  - Important
  - Useful
  - Noteworthy
  - Newsworthy
- It especially does not mean that H_a is true
- It means: $P(E|H_0) \leq \alpha$
- Choice of α (significance level) depends on:
  - Field of study
  - Sample size
  - Cost of a false positive
- Important: choose α before conducting the experiment!

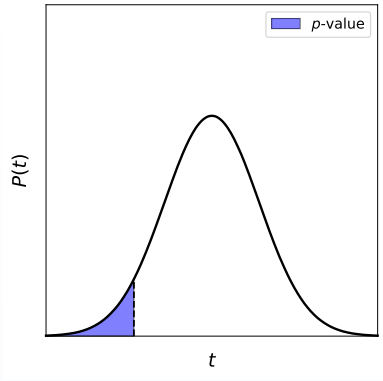
Critical values

- Rejection region: values of test statistic that lead to rejection of H_0
- Nonrejection region: values of test statistic that leads to non-rejection of H_0
- Critical value: the value of the test statistic that separates the rejection, nonrejection regions
  - Left-tailed: value where cdf equals α
  - Right-tailed: value where cdf equals $1 - \alpha$
  - Two-tailed: values where cdf equals $\alpha/2$, $1 - \alpha/2$
- Compare critical value for chosen α to your test statistic, t
  - If $t \leq t_{\alpha}$ (right-tail: $t \geq t_{1-\alpha}$; two-tail: $|t| \geq t_{1-\alpha/2}$), reject H_0
  - If $t > t_{\alpha}$ (right-tail: $t < t_{1-\alpha}$; two-tail: $|t| < t_{1-\alpha/2}$), do not reject H_0



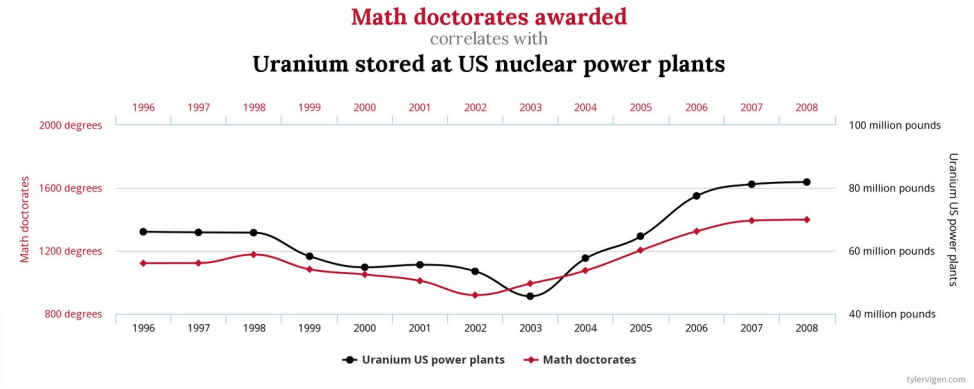
p-values

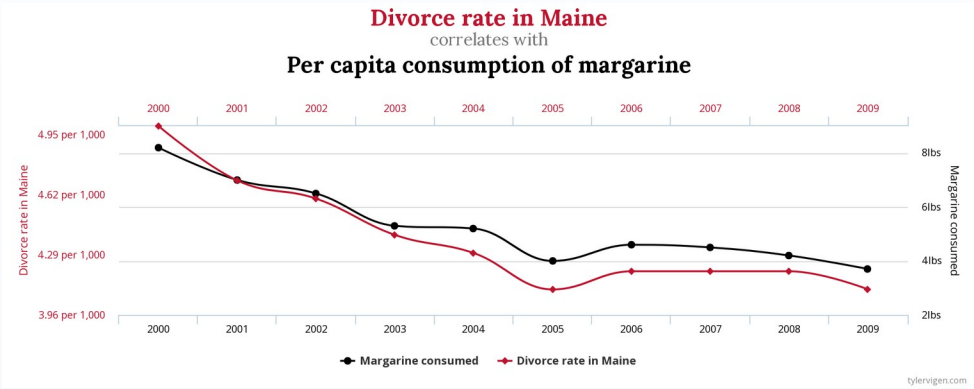
- p-value: the probability of seeing observations at least as extreme as ours, assuming that the null hypothesis is correct [$P(E|H_0)$]
- Instead of critical value, we can directly compare p-value and α
- What the p-value tells us:
  - How likely it is to see the observations, given that H_0 is correct
  - Whether or not we can reject H_0
- What the p-value **absolutely does not** tell us:
  - The probability that H_0 is correct [$P(H_0)$]
  - The probability that H_a is correct, given the evidence [$P(H_a|E)$]
  - Whether or not H_0 is false



p-hacking

- What happens if we test lots of hypotheses?
  - We will probably find some “significant” relationships or results
  - Remember: *p*-value of 0.05 → there’s a 1 in 20 chance of seeing the same observations by chance!
- *p*-hacking: performing lots of analyses, selecting only the “best” ones
  - Instead, split data: one for developing hypothesis, one for testing significance
  - Or: develop hypothesis, collect new data
- Important to remember: the *p*-value should not be the only evidence!





tylervigen.com/spurious-correlations

Summary

- Hypothesis testing: the formal process of drawing conclusions from data
  - Null hypothesis: there is no effect/difference
  - Alternative hypothesis: there is some effect/difference
- Always remember:
  - We **are not proving** that the null hypothesis is false
  - Nor are we proving that the alternative hypothesis is true
- Instead:
  - Determining if the evidence is good enough to reject the null hypothesis at a particular level of significance

Additional Resources

- Illowsky and Dean, Chapter 9
- Caswell, Chapter 15
- Weiss, Chapter 9
- Statistical inference: definition, methods & example [Jim Frost]
- Understanding significance levels in statistics [Jim Frost]
- Idea behind hypothesis testing [Khan Academy]
- The method that can “prove” almost anything [TED-Ed]
- Hack your way to scientific glory [FiveThirtyEight]